DOCUMENT RESUME

ED 403 276 TM 025 950

AUTHOR Roberts, James S.; Laughlin, James E.
TITLE The Graded Unfolding Model: A Unidimensional Item Response Model for Unfolding Graded Responses.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-16
PUB DATE Apr 96
NOTE 85p.
PUB TYPE Reports - Evaluative/Feasibility. (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Attitude Measures; Estimation (Mathematics); *Item Response Theory; *Mathematical Models; *Maximum Likelihood Statistics; Psychometrics; Simulation; Student Attitudes; *Test Items
IDENTIFIERS Binary Data Analysis; *Graded Response Model; *Unfolding Technique

ABSTRACT

Binary or graded disagree-agree responses to attitude items are often collected for the purpose of attitude measurement. Although such data are sometimes analyzed with cumulative measurement models, recent investigations suggest that unfolding models are more appropriate (J. S. Roberts, 1995; W. H. Van Schuur and H. A. L. Kiers, 1994). Advances in item response theory (IRT) have led to the development of several parametric unfolding models for binary data (D. Andrich, 1988; Andrich and G. Luo, 1993; H. Hoijtink, 1991), but IRT models for unfolding graded responses have not been addressed in the psychometric literature. A parametric IRT model for unfolding either binary or graded responses was developed in this study. The model, called the graded unfolding model (GUM) is a generalization of the hyperbolic cosine model for binary data of Andrich and Luo (1993). A joint maximum likelihood procedure was implemented to estimate GUM parameters, and a subsequent recovery simulation showed that reasonably accurate estimates could be obtained with minimal data demands (e.g. as few as 100 subjects and 15 to 20 6-category items). The applicability of the GUM to common attitude testing situations was illustrated with real data on student attitudes toward capital punishment. Appendixes discuss developing initial parameter values and the illustrative example. (Contains 1 table, 15 figures, 6 appendix figures, and 42 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-96-16

ED 403 276

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

THE GRADED UNFOLDING MODEL: A UNIDIMENSIONAL ITEM RESPONSE MODEL FOR UNFOLDING GRADED RESPONSES

James S. Roberts
James E. Laughlin

Educational Testing Service
Princeton, New Jersey
April 1996

TM025950

BEST COPY AVAILABLE

The Graded Unfolding Model: A Unidimensional Item Response Model for Unfolding Graded Responses

James S. Roberts

Educational Testing Service

James E. Laughlin

University of South Carolina

Copyright © 1996. Educational Testing Service. All rights reserved.

Abstract

Binary or graded disagree-agree responses to attitude items are often collected for the purpose of attitude measurement. Although such data are sometimes analyzed with cumulative measurement models, recent investigations suggest that unfolding models are more appropriate (Roberts, 1995; Van Schuur & Kiers, 1994). Advances in item response theory (IRT) have led to the development of several parametric unfolding models for binary data (Andrich, 1988; Andrich & Luo, 1993; Hoijtink, 1991), but IRT models for unfolding graded responses have not been addressed in the psychometric literature. A parametric IRT model for unfolding either binary or graded responses was developed in this study. The model, called the graded unfolding model (GUM), is a generalization of Andrich & Luo's (1993) hyperbolic cosine model for binary data. A joint maximum likelihood procedure was implemented to estimate GUM parameters, and a subsequent recovery simulation showed that reasonably accurate estimates could be obtained with minimal data demands (e.g., as few as 100 subjects and 15 to 20 6-category items). The applicability of the GUM to common attitude testing situations was illustrated with real data on student attitudes toward capital punishment. *Index terms: attitude measurement, graded unfolding model, hyperbolic cosine model, ideal point process, item response theory, Likert scale, Thurstone scale, unidimensional scaling, unfolding model.*

The use of disagree-agree responses to measure individual attitudes has a long history within psychology. The practice can be traced to Thurstone's (1928, 1931) classic approach in which individuals provide binary disagree-agree responses to a set of pre-scaled attitude statements, and then the responses are used to develop estimates of each individual's attitude. Use of disagree-agree responses has been bolstered further by the popularity of Likert's (1932) attitude measurement procedure in which individuals respond to a set of attitude items using a graded scale of agreement (e.g., "strongly disagree", "disagree", "agree", "strongly agree"), and the polytomous responses are subsequently used to develop a summated attitude score for each person.

Methods used to analyze disagree-agree responses[1] to attitude items are generally consistent with one of two perspectives of the response process. The first perspective suggests that disagree-agree responses result from an ideal point process (Coombs, 1964) in which an individual agrees with an attitude item to the extent that the item content satisfactorily represents the individual's own opinion. From this viewpoint, disagree-agree responses are best analyzed with some type of unfolding model that implements a single-peaked response function. Thurstone's (1928, 1931) attitude measurement procedure is an example of the ideal point perspective in that individual's are assumed to have attitudes that are similar to the items they endorse. More contemporary examples include parametric item response models such as the

[1] The term "disagree-agree responses" is used here in a generic sense which encompasses both binary (e.g. "disagree" and "agree") responses and graded (e.g., "strongly disagree", "disagree", "agree", "strongly agree") responses. The number of response categories associated with a disagree-agree response should be determined from the context in which the term is used unless such information is explicitly stated in the text.

squared simple logistic model (Andrich, 1988), the PARELLA model (Hoijtink, 1990, 1991), and the hyperbolic cosine model (Andrich & Luo, 1993). Nonparametric item response models have also been proposed for data that result from an ideal point response process (Cliff, Collins, Zatkin, Gallipeau & McCormick, 1988; van Schuur, 1984), but these nonparametric models do not yield measures which are invariant to either the items or the persons sampled in a particular application. The parametric item response models, in contrast, can yield invariant measures provided that the model in question is appropriate for the data (Hoijtink, 1990).

A second perspec.ive implies that disagree-agree responses are the result of a dominance process (Coombs, 1964) in which an individual agrees with a positively worded attitude statement to the extent that the individual's own opinion is more positive than the sentiment expressed in the statement. (Conversely, an individual agrees with a negatively worded statement to extent that the individual's opinion is more negative than the sentiment reflected by the statement.) According to this viewpoint, disagree-agree responses are best analyzed by some type of cumulative model that implements a monotonic response function. The Likert (1932) and Guttman (1950) approaches to attitude measurement illustrate classical techniques that are consistent with the dominance perspective. The application of cumulative item response models [e.g., the one-parameter (Rasch, 1960), two-parameter (Birnbaum, 1968) and three-parameter (Lord, 1980) logistic models, the graded response model (Samejima, 1969), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), and the generalized partial credit model (Muraki, 1992)] to disagree-agree responses provides a more contemporary illustration of the dominance perspective.

Several researchers have recently argued that disagree-agree responses are generally more

consistent with an ideal point perspective rather than a dominance perspective (Roberts, 1995; Roberts, Laughlin & Wedell, 1996; van Schuur & Kiers, 1994). This argument implies that attitude measures based on disagree-agree responses are more appropriately developed from unfolding models rather than cumulative models. Moreover, Roberts (1995) has shown that cumulative models can yield attitude measures which are nonmonotonically related to the latent trait when such models are applied to responses from an ideal point process. Specifically, individuals with the most extreme attitudes may receive scores that are indicative of more moderate attitudinal positions.

Although unfolding models appear most appropriate for disagree-agree data, the application of unfolding item response models to attitude measurement remains somewhat problematic because such models allow only for binary disagree-agree responses. Some researchers, however, have documented gains in precision when polytomous responses, as opposed to binary responses, are used to derive measurements from cumulative item response models (Bock, 1972; Donoghue, 1994; Thissen, 1976), and Roberts (1995) has suggested that similar benefits can be achieved with polytomous unfolding models. Furthermore, psychologists often collect graded responses to attitude statements in the traditional Likert (1932) fashion, and these practitioners would be forced to dichotomize their data and risk losing valuable information before using any of the contemporary unfolding item response models. What is needed is an unfolding item response model that incorporates a graded scale of agreement (i.e., graded disagree-agree responses).

In the following pages, an unfolding item response model is proposed for disagree-agree responses that result from either a binary or graded scale. The model, referred to as the graded unfolding model (GUM), is a generalization of Andrich and Luo's (1993) hyperbolic cosine model

for binary data. The GUM incorporates a single peaked response function, and therefore, it is applicable in situations where responses are presumably generated from an ideal point process. Additionally, because the GUM allows for graded responses, there is no need to dichotomize polytomous data prior to estimating model parameters, and thus, there is no corresponding loss in the precision of person estimates.

The Graded Unfolding Model (GUM)

The GUM is developed from a series of five basic premises about the response process. The first premise is that when an individual is asked to express his or her agreement with an attitude statement, then the individual tends to agree with the item to the extent that it is located close to his or her own position on a unidimensional latent attitude continuum. In this context, the degree to which the sentiment of an item reflects the opinion of an individual is given by the proximity of the individual to the item on the attitude continuum. If we let δ_i denote the position of the *ith* item on the continuum and let θ_n denote the location of the *nth* individual on the continuum, then the individual is more likely to agree with the item to the extent that the distance between θ_n and δ_i approaches zero. This is simply a restatement of the fundamental characteristic of an ideal point process (Coombs, 1964).

A second premise of the GUM is that an individual may respond in a given response category for either of two distinct reasons. As an example, consider an individual with a neutral attitude toward capital punishment. This individual might strongly disagree with an item that portrays the practice of capital punishment in either a very negative or very positive way. If the item is located far below the individual's position on the attitude continuum (i.e., the item's content is much more

negative than the individual's attitude), then we would say that the individual "strongly disagrees from above" the item. In contrast, if the item is located far above the individual's position (i.e., the item's content is much more positive than the individual's attitude), then we would say that the individual "strongly disagrees from below" the item. Hence, there are two possible subjective responses, "strongly disagree from above" and "strongly disagree from below", associated with the single observable response of "strongly disagree". Similarly, the GUM postulates two subjective responses for each observable response on a rating scale.

The third premise behind the GUM is that subjective responses to attitude statements follow a cumulative item response model (e.g., Andrich & Luo, 1993). In this paper, we will assume that subjective responses follow Andrich's (1978) rating scale model, but other cumulative models could also be used.[2] The rating scale model for subjective responses can be defined as:

$$Pr[Y_i = y \mid \theta_n] = \frac{\exp[y(\theta_n - \delta_i) - \sum_{j=0}^{y} \tau_j]}{\sum_{w=0}^{M} [\exp[w(\theta_n - \delta_i) - \sum_{j=0}^{w} \tau_j]}, \quad (1)$$

where:

Y_i = a subjective response to attitude statement i;

$y = 0, 1, 2, \ldots, M$; $y = 0$ corresponds to the strongest level of disagreement from below the item whereas $y = M$ corresponds to the strongest level of disagreement from above the item (see Figure 1),

[2] The rating scale model seemed particularly appropriate given that the same response scale is typically used for all items on a traditional Likert or Thurstone attitude questionnaire.

θ_n = the location of individual *n* on the attitude continuum,

δ_i = the location of attitude statement *i* on the attitude continuum,

τ_j = the relative location of the *jth* subjective response category threshold on the attitude continuum (relative to a given item),

M = the number of subjective response categories minus 1.

The notation used on the left side of equation 1 implicitly conditions on the item parameters δ_i and τ_j. The model is illustrated in Figure 1 for a hypothetical item with four observable response categories: "strongly disagree", "disagree", "agree", "strongly agree". The abscissa of Figure 1 represents the attitude continuum, and it is scaled in units of signed distance between an individual's attitude position and the location of the item (i.e., $\theta_n - \delta_i$). The ordinate indexes the probability that an individual's subjective response will fall in one of the 8 possible subjective response categories. (There are 8 subjective response categories and associated probability curves due to the fact that an individual may respond in any of the 4 observable response categories because his or her attitudinal position is either above or below the location of the item.) The 7 vertical lines designate the locations where successive subjective response category probability curves intersect. These locations are the subjective response category thresholds. In this example, the 7 subjective response category thresholds are successively ordered on the latent continuum. Therefore, these thresholds divide the latent continuum into 8 intervals in which a different subjective response is most likely. The most likely subjective response within each interval is labeled in the figure.

Insert Figure 1 About Here

Equation 1 defines an item response model at a subjective response level. However, the model must ultimately be defined in terms of the observable response categories associated with the graded agreement scale. Recall that each observable response category is associated with two possible subjective responses (i.e., one from below the item and one from above the item). Moreover, the two subjective responses corresponding to a given observable response category are mutually exclusive. Therefore, the probability that an individual will respond using a particular observable category is simply the sum of the probabilities associated with the two corresponding subjective responses:

$$Pr[Z_i = z \mid \theta_n] = Pr[Y_i = z \mid \theta_n] + Pr[Y_i = (M-z) \mid \theta_n], \quad (2)$$

where:

Z_i = an observable response to attitude statement i,

z = 0, 1, 2, ..., C ; $z = 0$ corresponds to the strongest level of disagreement and $z = C$ refers to the strongest level of agreement,

C = the number of observable response categories minus 1. Note that $M = 2*C + 1$.

The fourth premise underlying the GUM is that subjective responses are locally independent across multiple items. Therefore, subjective responses are presumed to be uncorrelated at a fixed ability level. This premise also implies that observable responses are locally independent due to the relationship between observable and subjective responses.

The fifth and final premise behind the GUM is that subjective category thresholds are symmetric about the point $(\theta_n - \delta_i) = 0$, which implies that $\tau_z = -\tau_{(M-z+1)}$ for all $z \neq 0$ and $\tau_{(C+1)} = 0$. At a conceptual level, this premise suggests that an individual is just as likely to agree with an item

located at either $-x$ units or $+x$ units from the individual's position on the attitude continuum. At an analytical level, this premise leads to the following identity:

$$\sum_{j=0}^{z} \tau_j = \sum_{j=0}^{M-z} \tau_j \quad . \tag{3}$$

Incorporating this identity into equation 2 yields the formal definition of the GUM:

$$Pr[Z_i = z \mid \theta_n] = \frac{\exp[z(\theta_n - \delta_i) - \sum_{j=0}^{z} \tau_j] + \exp[(M-z)(\theta_n - \delta_i) - \sum_{j=0}^{z} \tau_j]}{\sum_{w=0}^{C} \left(\exp[w(\theta_n - \delta_i) - \sum_{j=0}^{w} \tau_j] + \exp[(M-w)(\theta_n - \delta_i) - \sum_{j=0}^{w} \tau_j] \right)} \quad . \tag{4}$$

Note that the parameterization used in equation 4 requires only a single constraint on item parameter values:

$$\sum_{i=1}^{I} \delta_i = 0 \tag{5}$$

along with the notational definition that τ_0 equals zero.[3]

Insert Figure 2 About Here

The GUM defines the observable response category probability curves associated with the *nth* individual's objective response to the *ith* item. Figure 2 displays these probability curves for the same hypothetical item referenced in Figure 1. As seen in Figure 2, there is one probability curve

---

[3] The value of τ_0 could arbitrarily be set equal to any constant without affecting any of the probabilities defined by Equation 4 (Muraki, 1992).

associated with each observable response available to the individual. Each of these probability curves is simply the sum of the two corresponding subjective response category probability curves previously shown in Figure 1. One should note that successive observable response category probability curves do not intersect at τ_1, τ_2, ..., τ_C, and therefore, the τ_j parameters lose their simple interpretation at the observable score level. In contrast, the substantive meaning of both θ_n and δ_i remains unchanged when moving from a subjective score level to an observable score level.

Insert Figure 3 About Here

The GUM is an unfolding model of the response process. This is easily seen by computing the expected value of an observable response for various values of $\theta_n - \delta_i$ using the probability function given in equation 4. Figure 3 portrays the expected value of an observable response for the same hypothetical item with 4 response categories. The categories are coded with the integers 0 to 3 where the codes correspond to the responses of "strongly disagree", "disagree", "agree" and "strongly agree", respectively. As seen in Figure 3, the item elicits greater levels of agreement as the distance between the individual and the item on the attitude continuum decreases.

Parameter Estimation

In this study, the characteristics of a joint maximum likelihood approach to parameter estimation will be described. This approach is based heavily on the (unconditional) procedures described by Andrich (1978), Masters (1982) and Wright and Masters (1982). Assuming that responses are independent across subjects and that the responses from any one subject are locally independent, then the likelihood function for the GUM may be written as:

$$L = \prod_{n=1}^{N} \prod_{i=1}^{I} Pr[Z_i = X_{ni} \mid \theta_n, \delta_i, \tau_j], \tag{6}$$

where X_{ni} is the *nth* individual's observed response to item *i*. The goal of the maximum likelihood procedure is to find the values of θ_n, δ_i, and τ_j that maximize the probability of obtaining the observed data, and hence, the values that maximize equation 6. In practice, the natural logarithm of likelihood equation is maximized instead of the likelihood function itself. The log-likelihood function is equal to:

$$\ln(L) = \sum_{n=1}^{N} \sum_{i=1}^{I} \left[\left(-\sum_{j=0}^{X_{ni}} \tau_j\right) + \ln\left(\left(\exp[X_{ni}(\theta_n - \delta_i)] + \exp[(M - X_{ni})(\theta_n - \delta_i)]\right) \right.\right.$$

$$\left.\left. - \ln\left(\sum_{w=0}^{C} \left(\exp\left[-\sum_{j=0}^{w} \tau_j\right]\right)\left(\exp[w(\theta_n - \delta_i)] + \exp[(M-w)(\theta_n - \delta_i)]\right) \right) \right]. \tag{7}$$

In our approach, the parameter values that maximize the log-likelihood function are determined in an iterative fashion following the logic of Wright and Masters (1982). First, the value of τ_1 that maximizes the log-likelihood function is determined while treating all other parameters as constants. This maximization process is subsequently repeated for each of the remaining τ_j parameters. The log-likelihood function is then maximized with respect to each of the δ_i parameters in succession, after which, the location constraint in equation 5 is imposed. Finally, the log-likelihood function is maximized with regard to each of the θ_n parameters. Maximizing the log-likelihood function with respect to every parameter constitutes one maximization cycle within the iterative procedure. Maximization cycles are repeatedly performed until the average change in δ_i and τ_j parameters is less than some arbitrarily small value (e.g., .005).

Within any cycle, maximization of the log-likelihood function with respect to a given parameter, Φ, is accomplished through the Newton-Raphson algorithm. The Newton-Raphson algorithm finds the value of Φ at which the partial derivative of the log-likelihood function with respect to Φ is zero. This root is determined iteratively, such that the value of Φ at iteration $q+1$ is equal to:

$$\Phi_{q+1} = \Phi_q - \frac{\left[\frac{\partial \ln(L)}{\partial \Phi}\right]_q}{\left[\frac{\partial^2 \ln(L)}{\partial \Phi^2}\right]_q}. \qquad (8)$$

Thus, the first and second order partial derivatives of the log-likelihood function with respect to Φ are required in order to estimate GUM parameters. These partial derivatives are more easily computed if we rewrite the log-likelihood function as:

$$\ln(L) = \sum_{n=1}^{N} \sum_{i=1}^{I} \left[-\sum_{j=0}^{C} U_{X_{ni}j}(\tau_j) + \ln\left(\exp[X_{ni}(\theta_n - \delta_i)] + \exp[(M - X_{ni})(\theta_n - \delta_i)] \right) \right.$$
$$\left. - \ln\left(\sum_{w=0}^{C} \left(\exp[-\sum_{j=0}^{C} U_{wj}(\tau_j)]\right) \left(\exp[w(\theta_n - \delta_i)] + \exp[(M-w)(\theta_n - \delta_i)]\right) \right) \right], \qquad (9)$$

where $U_{X_{ni}j}$ and U_{wj} are dummy variables that are equal to 1 when $j \leq X_{ni}$ or $j \leq w$, respectively, and are equal to 0 otherwise. The first order partial derivative with respect to θ_n is then:

$$\frac{\partial \ln(L)}{\partial \theta_n} = \sum_{i=1}^{I} \left(\frac{b_{ni}(X_{ni}) + c_{ni}(M-X_{ni})}{b_{ni} + c_{ni}} - \frac{\sum_{w=0}^{C} d_w (e_{wni} w + f_{wni}(M-w))}{\gamma_{ni}} \right)$$

$$= \sum_{i=1}^{I} \left(\frac{\frac{[b_{ni}(X_{ni}) + c_{ni}(M-X_{ni})]a_{ni}}{\gamma_{ni}}}{Pr[Z_i = X_{ni}]} - \frac{\sum_{w=0}^{C} d_w (e_{wni} w + f_{wni}(M-w))}{\gamma_{ni}} \right)$$

$$= \sum_{i=1}^{I} \left[\frac{Pr[Y_i = X_{ni} \mid \theta_n](X_{ni}) + Pr[Y_i = M-X_{ni} \mid \theta](M-X_{ni})}{Pr[Z_i = X_{ni}]} \right.$$

$$\left. - \left(\sum_{w=0}^{C} Pr[Y_i = w \mid \theta_n](w) + Pr[Y_i = M-w \mid \theta](M-w) \right) \right] \qquad (10)$$

$$= \sum_{i=1}^{I} \left(E[Y_i \mid \theta_n, X_{ni}] - E[Y_i \mid \theta_n] \right),$$

where:

$$a_{ni} = \exp[- \sum_{j=0}^{C} U_{X_{ni}j} \tau_j], \qquad (11)$$

$$b_{ni} = \exp[X_{ni}(\theta_n - \delta_i)], \qquad (12)$$

$$c_{ni} = \exp[(M-X_{ni})(\theta_n - \delta_i)], \qquad (13)$$

$$d_w = \exp[- \sum_{j=0}^{C} U_{wj} \tau_j], \qquad (14)$$

$$e_{wni} = \exp[w(\theta_n - \delta_i)], \qquad (15)$$

$$f_{wni} = \exp[(M-w)(\theta_n - \delta_i)], \qquad (16)$$

and

$$\gamma_{ni} = \sum_{w=0}^{C} d_w (e_{wni} + f_{wni}) . \qquad (17)$$

Note that in equation 10, $E[Y_i \mid \theta_n]$ is the expectation of the *nth* individual's subjective response to item *i*, and $E[Y_i \mid \theta_n, X_{ni}]$ is the conditional expectation of the *nth* individual's subjective response given that individual's observed response. Similarly, the partial derivatives with respect to δ_i and τ_j are:

$$\frac{\partial \ln(L)}{\partial \delta_i} = -\sum_{n=1}^{N} \left(E[Y_i \mid \theta_n, X_{ni}] - E[Y_i \mid \theta_n] \right), \qquad (18)$$

and

$$\frac{\partial \ln(L)}{\partial \tau_j} = -\sum_{n=1}^{N} \sum_{i=1}^{I} \left(U_{X_{ni}j} + \frac{\sum_{w=0}^{C} d_w (-U_{wj})(e_{wni} + f_{wni})}{\gamma_{ni}} \right) \qquad (19)$$

$$= -\sum_{n=1}^{N} \sum_{i=1}^{I} \left(U_{X_{ni}j} - \sum_{w=0}^{C} (U_{wj}) \, Pr[Z_i = w] \right) .$$

The second order partial derivatives of the log-likelihood equation are equal to:

$$\frac{\partial^2 \ln(L)}{\partial \theta_n^2} = \sum_{i=1}^{I} \left[\left(\frac{(b_{ni} + c_{ni})(b_{ni} X_{ni}^2 + c_{ni}(M - X_{ni})^2) - (b_{ni} X_{ni} + c_{ni}(M - X_{ni}))^2}{(b_{ni} + c_{ni})^2} \right) \right.$$

$$\left. - \left(\frac{(\sum_{w=0}^{C} d_w (e_{wni} w^2 + f_{wni} (M-w)^2)) \gamma_{ni} - (\sum_{w=0}^{C} d_w (e_{wni} w + f_{wni} (M-w)))^2}{\gamma_{ni}^2} \right) \right] \qquad \textbf{(20)}$$

$$= \sum_{i=1}^{I} \left[\left(E[Y_i^2 \mid \theta_n, X_{ni}] - E[Y_i \mid \theta_n, X_{ni}]^2 \right) - \left(E[Y_i^2 \mid \theta_n] - E[Y_i \mid \theta_n]^2 \right) \right]$$

$$= \sum_{i=1}^{I} \sigma^2_{Y_i \mid \theta_n, X_{ni}} - \sigma^2_{Y_i \mid \theta_n},$$

$$\frac{\partial^2 \ln(L)}{\partial \delta_i^2} = \sum_{n=1}^{N} \sigma^2_{Y_i \mid \theta_n, X_{ni}} - \sigma^2_{Y_i \mid \theta_n}, \qquad \textbf{(21)}$$

and

$$\frac{\partial^2 \ln(L)}{\partial \tau_j^2} = \sum_{n=1}^{N} \sum_{i=1}^{I} \left[\left(\sum_{w=0}^{C} U_{wj} Pr[Z_i = w] \right)^2 - \left(\sum_{w=0}^{C} U_{wj} Pr[Z_i = w] \right) \right], \qquad \textbf{(22)}$$

where:

$E[\, Y_i^2 \mid \theta_n \,]$ = the expectation of an individual's squared subjective response to item i,

$E[\, Y_i^2 \mid \theta_n, X_{ni} \,]$ = the conditional expectation of an individual's squared subjective response to item i given the individual's observable response,

$\sigma^2_{Y_i \mid \theta_n}$ = the variance of an individual's subjective response to item i,

and

$\sigma^2_{Y_i|\theta_n, X_{ni}}$ = the conditional variance of an individual's subjective response to item *i* given the individual's observable response to item *i*.

Local Maxima of the Log-Likelihood Function

The log-likelihood function of the GUM with respect to either persons or items need not be single peaked. In practice, we have found that when the log-likelihood function contains local maxima with respect to δ_i, then these maxima are usually located relatively far from the point where the global maximum is attained. Furthermore, analyses of both simulated and real data have indicated that the values of the log-likelihood function at these local maxima are usually much smaller than the global maximum value. Under these circumstances, the Newton-Raphson algorithm appears to perform adequately provided that judicious initial values for δ_i are used. (A strategy for developing initial estimates is described in Appendix A.)

Our past experience with the model has shown that when local maxima occur in the log-likelihood function with respect to θ_n, then they may occur at points on the latent continuum which are relatively close to the location of the global maximum. Moreover, the value of the log-likelihood function at these local maxima can be quite similar to the global maximum value. In these situations, the Newton-Raphson procedure often falters. Therefore, the maximization algorithm for person parameters includes a grid search that is conducted at various points in the Newton-Raphson process in an effort to prevent locally optimal solutions (see Roberts, 1995).

Extreme Response Patterns

The joint maximum likelihood procedure will yield infinite attitude estimates for individuals who consistently use the most extreme disagree category (i.e., for individuals who receive a score of 0 on each item). Moreover, this extreme response pattern contains no information about the

direction of the individual's attitude. For example, an individual with this response pattern may possess an attitude that is extremely negative or extremely positive relative to the items under consideration. For these reasons, individuals who exhibit this extreme response pattern must be discarded from the estimation process. Additionally, experience has shown that erratic attitude estimates may arise in cases where an individual has not endorsed any item to at least a slight extent. Therefore, those individuals who fail to use any response categories that reflect agreement are also discarded from the estimation procedure.

Standard Errors of GUM Parameter Estimates

Approximate standard errors of the GUM parameters can be derived from the second order partial derivatives of the log-likelihood function. For example, if item parameters are treated as known, then approximate standard errors for attitude estimates can be computed as follows:

$$
\begin{aligned}
\hat{\sigma}_{\hat{\theta}_n} &= \left\{ - E\left[\frac{\partial^2 \ln(L)}{\partial \theta_n^2} \right] \right\}^{-1/2} \\
&= \left\{ - \sum_{i=1}^{I} \sum_{z=0}^{C} Pr[Z_i = z] \left[\left(\frac{(\tilde{b}_{ni} + \tilde{c}_{ni})(\tilde{b}_{ni} z^2 + \tilde{c}_{ni}(M-z)^2) - (\tilde{b}_{ni} z + \tilde{c}_{ni}(M-z))^2}{(\tilde{b}_{ni} + \tilde{c}_{ni})^2} \right) \right. \right. \\
&\quad \left. \left. - \left(\frac{(\sum_{w=0}^{C} d_w (e_{wni} w^2 + f_{wni} (M-w)^2)) \gamma_{ni} - (\sum_{w=0}^{C} d_w (e_{wni} w + f_{wni} (M-w)))^2}{\gamma_{ni}^2} \right) \right] \right\}^{-1/2} \\
&= \left\{ - \sum_{i=1}^{I} \left(\sum_{z=0}^{C} \left[Pr[Z_i = z] \, \sigma^2_{Y_i | \theta_n, z} \right] \right) - \sigma^2_{Y_i | \theta_n} \right\}^{-\frac{1}{2}},
\end{aligned}
\tag{23}
$$

where:

$\tilde{b}_{ni} = \exp[\, z\,(\theta_n - \delta_i)\,]$,

$\tilde{c}_{ni} = \exp[(M - z)\,(\theta_n - \delta_i)]$,

and

$\sigma^2_{Y_i|\theta_n, z}$ = the conditional variance of an individual's subjective response given that $Z_i = z$.

Note that $\partial^2 \ln(L)/\partial\theta_n^2$ involves observed data, and thus, the expected values of these terms are obtained numerically (i.e., calculated over $z = 0, 1, \ldots, C$) in equation 23. Furthermore, the nominal values of θ_n in this equation will necessarily be replaced by estimated θ_n values in practice.

The standard errors in equation 23 will generally be too small because the uncertainty in item parameter estimates is ignored. However, simulation results suggest that the degree of underestimation will be minor in many attitude measurement situations. These results will be described more fully in a subsequent section.

Insert Figure 4 About Here

The standard error of a θ_n estimate based on a single individual's response to a single item is plotted in Figure 4. The plot is based on a hypothetical item with six response categories under the assumption that successive τ_j values are equally distant from each other. Each curve in Figure 4 corresponds to an alternative value of ψ, where ψ is the distance between successive thresholds. The function is symmetric about the origin, and it becomes infinitely large whenever $|\theta_n - \delta_i|$ is equal to 0 or is infinitely large itself. The global minimum of the function increases as ψ increases, and the points (or intervals) on the $\theta_n - \delta_i$ axis at which this minimum occurs grow more distant

from the origin. Moreover, the range of $|\theta_n - \delta_i|$ values which yield standard errors near the global minimum gets larger as ψ increases. Thus, standard errors become larger, yet more stable, as ψ increases.

Approximate variances and covariances of item parameter estimates can be obtained by treating person parameters as known. The approximations are given by matrix Σ where:

$$\Sigma = \begin{bmatrix} \xi_{\delta\delta} & \xi_{\delta\tau} \\ \xi_{\tau\delta} & \xi_{\tau\tau} \end{bmatrix}^{-1} \cdot \cdot \qquad \textbf{(24)}$$

Matrix Σ is an $(I + C)$ x $(I + C)$ matrix of variances and covariances formed by juxtaposing four submatrices. Standard errors of item parameters are obtained by taking the square roots of the diagonal elements of Σ. Submatrix $\xi_{\delta\delta}$ is an I x I diagonal matrix where the *ith* diagonal element is given by:

$$\begin{aligned} \xi_{\delta_i\delta_i} &= \left\{ - E\left[\frac{\partial^2 \ln(L)}{\partial \delta_n^2} \right] \right\} \\ &= \left\{ - \sum_{n=1}^{N} \left(\sum_{z=0}^{C} \left[Pr[Z_i = z \,] \, \sigma^2_{Y_i|\theta_n, z} \right] \right) - \sigma^2_{Y_i|\theta_n} \right\}. \end{aligned} \qquad \textbf{(25)}$$

The expectation in equation 25 is derived numerically across all possible values of z. Submatrix $\xi_{\tau\tau}$ is a C x C symmetric matrix in which the (j, j') element is given by:

$$\begin{aligned} \xi_{\tau_j\tau_{j'}} &= \left[- \frac{\partial^2 \ln(L)}{\partial \tau_j \, \partial \tau_{j'}} \right] \\ &= \sum_{n=1}^{N} \sum_{i=1}^{I} \sum_{z=0}^{C} \left[U_{zj} \, Pr[Z_i = z] \left(- U_{zj'} + \sum_{v=0}^{C} U_{vj'} \, Pr[Z_i = v \,] \right) \right], \end{aligned} \qquad \textbf{(26)}$$

where U_{zj}, $U_{zj'}$ and $U_{vj'}$ are dummy variables which are equal to 1 whenever $z \leq j$, $z \leq j'$ or $v \leq j'$, respectively, and are otherwise equal to zero. No expectation is required in equation 26 because it contains no observed data. Submatrix $\xi_{\delta\tau}$ is an I x C matrix in which the (i, j) element is equal to:

$$\begin{aligned}
\xi_{\delta_i \tau_j} &= \left[- \frac{\partial^2 \ln(L)}{\partial \delta_i \partial \tau_j} \right] \\
&= \sum_{n=1}^{N} \sum_{z=0}^{C} U_{zj} \left[\left(\frac{\tilde{a}_z [z \tilde{b}_{ni} + (M-z) \tilde{c}_{ni}]}{\gamma_{ni}} \right) - Pr[Z_i = z] \left(\frac{\sum_{w=0}^{C} d_w (w e_{wni} + (M-w) f_{wni})}{\gamma_{ni}} \right) \right] \\
&= \sum_{n=1}^{N} \sum_{z=0}^{C} U_{zj} \left[\left(\frac{\tilde{a}_z [z \tilde{b}_{ni} + (M-z) \tilde{c}_{ni}]}{\gamma_{ni}} \right) - Pr[Z_i = z] \left(E[Y_i | \theta_n] \right) \right] \\
&= \sum_{n=1}^{N} \sum_{z=0}^{C} U_{zj} Pr[Z_i = z] \left(E[Y_i | \theta_n, z] - E[Y_i | \theta_n] \right),
\end{aligned} \tag{27}$$

where:

$\tilde{a}_z = \exp [- \sum_{j=0}^{C} U_{zj} \tau_j]$,

and

$E[Y_i | \theta_n, z]$ = the conditional expectation of an individual's subjective response to item i given an observable response of z.

Equation 27 contains no observed data, and thus, no expectation is required. The fourth and final submatrix, $\xi_{\tau\delta}$, is simply the transpose of submatrix $\xi_{\delta\tau}$. Note that the nominal item parameter values given in equations 25 through 27 are necessarily replaced by parameter estimates in practice.

The variances and covariances defined by equation 24 will generally be too small because the uncertainty in person estimates is ignored. Nonetheless, simulation results suggest that the

approximations will often be adequate, and their accuracy will be discussed later in this report.

Recovery of Simulated Parameters

Method

The recovery of GUM parameters by means of the joint maximum likelihood technique was simulated under 30 alternative conditions. Conditions were derived by factorially combining five levels of sample size with six levels of test length. The number of subjects analyzed in these conditions was either 100, 200, 500, 1000 or 2000, and responses to either 5, 10, 15, 20, 25 or 30 attitude items were generated. Items were always located at equally distant positions on the latent continuum and always ranged from -4.25 to 4.25, regardless of the number of items studied. The person parameter estimates in each condition were randomly sampled from a normal distribution with $\mu=0$ and $\sigma=2.0$. The threshold parameters were equally distant, and the interthreshold distance was fixed at .4. The characteristics of the nominal item and person parameters used in this simulation were similar to those used in previous studies of estimation accuracy with unfolding IRT models (Andrich, 1988; Andrich & Luo, 1993). The observable response simulated for each person-item combination was on a 6-point scale (e.g., "strongly disagree", "disagree", "slightly disagree", "slightly agree", "agree" and "strongly agree"). Six response probabilities were computed with equation 4 and subsequently used to divide a probability interval (i.e., a closed interval between 0 and 1) into 6 mutually exclusive and exhaustive segments, where each segment corresponded to a particular observable response category. A random number was then generated from a uniform probability distribution, and the simulated response was that response associated with the probability segment in which the

random number fell. After an observable response to each item had been generated for all subjects, the data were used to estimate GUM parameters. The process of generating data and subsequently estimating parameters was replicated 30 times in each condition, and the nominal values of all parameters remained constant across replications.

Measures of Estimation Accuracy

Four measures of estimation accuracy were investigated, and each measure had previously been used in at least one study of accuracy in IRT parameter estimation (e.g., Andrich, 1988; Andrich & Luo, 1993; Hulin, Lissak & Drasgow, 1982; Kim, Cohen, Baker, Subkoviak & Leonard, 1994; Seong, 1990; Yen, 1987). The Root Mean Squared Error (RMSE) was the first of these measures. The RMSE provided an index of the average unsigned discrepancy between a set of nominal parameters and a corresponding set of estimates. The RMSE was calculated across all the parameters of a given type in any single replication. For example, the RMSE of attitude location estimates from a particular replication was computed as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (\hat{\theta}_n - \theta_n)^2}{N}}, \qquad (28)$$

where:

θ_n = the nominal attitude location for the *nth* subject,

$\hat{\theta}_n$ = the estimated attitude location for the *nth* subject,

and

N = the number of subjects in the sample.

Similar quantities were computed for the item location parameters and the threshold parameters in

a given replication. The RMSE for θ_n estimates can be algebraically decomposed to show that:

$$RMSE = \sqrt{S_{\hat{\theta}}^2 + S_{\theta}^2 - 2(S_{\hat{\theta}\theta}) + (\bar{X}_{\hat{\theta}} - \bar{X}_{\theta})^2}\,, \qquad (29)$$

where:

$S_{\hat{\theta}}^2$ = the sample variance of the N estimated θ values,

S_{θ}^2 = the sample variance of the N nominal θ values,

$S_{\hat{\theta}\theta}$ = the sample covariance between the estimated and nominal θ values,

$\bar{X}_{\hat{\theta}}$ = the average of the N estimated θ values,

and

$\bar{X}_{\theta}$ = the average of the N nominal θ values.[4]

The RMSE for δ_i and τ_j estimates can be decomposed in an analogous fashion. Equation 29 illustrates that the RMSE is sensitive to three types of discrepancies between the nominal and estimated parameter distributions. First, it depends on the degree of covariation between nominal and estimated parameter distributions. Relatively large RMSE values are expected when the linear relationship between estimated and nominal parameters is weak. Second the RMSE depends on the degree to which the variance of the estimated parameter distribution matches that for the nominal distribution. It will increase as the variances of the two distributions become more discrepant. Lastly, the RMSE depends on the extent to which the mean of the estimated parameter distribution matches that for the nominal parameter distribution. It will increase as the absolute difference between the two means grows larger.

---

[4] The terms "sample variance" and "sample covariance" are used to designate the biased forms of these statistics where N is used in the denominator rather than N-1.

The remaining accuracy measures were used to individually evaluate the three discrepancies assessed by the RMSE. For example, a Pearson correlation between estimated and nominal parameters was computed on each replication in order to index the degree of covariation between distributions. Similarly, the ratio of estimated parameter sample variance to nominal parameter sample variance was calculated to determine how well the variances of the two distributions matched. Lastly, the mean difference between estimated and nominal parameters was computed, and the absolute value of this mean difference was used as an estimate of the location difference between the two distributions.

Measures of Bias

In addition to the measures of accuracy described above, four measures of statistical bias were examined. The first of these measures, the Root Mean Square Bias (RMSB), provided a global measure of bias across all parameters of a given type. The RMSB for θ_n estimates was defined as:

$$RMSB = \sqrt{\frac{\sum_{n=1}^{N} (E[\hat{\theta}_n] - \theta_n)^2}{N}} \qquad (30)$$

$$= \sqrt{S^2_{E[\hat{\theta}_n]} + S^2_{\theta} - 2(S_{E[\hat{\theta}_n]\theta}) + (\bar{X}_{E[\hat{\theta}_n]} - \bar{X}_{\theta})^2},$$

where:

$E[\hat{\theta}_n]$ = the expected value of $\hat{\theta}$ for the *nth* individual,

$S^2_{E[\hat{\theta}_n]}$ = the sample variance of $E[\hat{\theta}_n]$ over the N individuals,

$S_{E[\hat{\theta}_n]\theta}$ = the sample covariance between the expected and nominal θ values over the N individuals,

and

$\bar{X}_{E[\hat{\theta}_n]}$ = the average of $E[\hat{\theta}_n]$ over the N individuals.

Analogous quantities were defined for the δ_i and τ_j estimates. The RMSB could not be calculated directly because the expected values of parameter estimates were unknown. Instead, it was approximated from the simulation data across the 30 replications in each cell of the design. For example, the values for $\bar{X}_{E[\hat{\theta}_n]}$, $S^2_{E[\hat{\theta}_n]}$, and $S_{E[\hat{\theta}_n]\theta}$ were approximated as follows:

$$\bar{X}_{E[\hat{\theta}_n]} \approx \frac{\sum_{n=1}^{N} \bar{\hat{\theta}}_n}{N} = \bar{X}_{\bar{\hat{\theta}}_n}, \tag{31}$$

$$S^2_{E[\hat{\theta}_n]} \approx S^2_{\bar{\hat{\theta}}} - \frac{\sum_{n=1}^{N} \frac{s^2_{\hat{\theta}_n}}{R}}{N}, \tag{32}$$

and

$$S_{E[\hat{\theta}_n]\,\theta} \approx S_{\bar{\hat{\theta}}\,\theta}, \tag{33}$$

where:

$\bar{\hat{\theta}}_n$ = the average θ estimate for the *nth* individual where the average is calculated across the R replications of the simulation,

$\bar{X}_{\bar{\hat{\theta}}_n}$ = the average of $\bar{\hat{\theta}}_n$ across the N individuals,

$S^2_{\bar{\hat{\theta}}}$ = the sample variance of $\bar{\hat{\theta}}_n$ across the N individuals,

$s^2_{\hat{\theta}_n}$ = the (unbiased) empirical estimate of the population error variance of $\hat{\theta}_n$ calculated across the R replications of the simulation [i.e., $s^2_{\hat{\theta}_n} = \sum_{r=1}^{R} (\hat{\theta}_n - \bar{\hat{\theta}}_n)^2 / (R-1)$],

$S_{\bar{\theta}\,\theta}$ = the sample covariance between $\bar{\hat{\theta}}_n$ and θ_n across the N individuals.

Similar approximations were developed for the δ_i and τ_j parameters.

Like the RMSE, the RMSB is sensitive to three forms of discrepancy between nominal and expected parameter distributions. It will be larger to the extent that the two distributions are not perfectly correlated, have different variances or have different means. Consequently, three additional bias measures were constructed in an effort to diagnose these particular discrepancies. For example, the correlation between expected and nominal θ_n distributions was approximated by:

$$r_{E[\hat{\theta}_n]\,\theta} = \frac{S_{E[\hat{\theta}_n]\,\theta}}{S_\theta\, S_{E[\hat{\theta}_n]}} \approx \frac{S_{\bar{\theta}\,\theta}}{S_\theta \sqrt{S^2_{\bar{\theta}} - \dfrac{\sum_{n=1}^{N} \dfrac{s^2_{\hat{\theta}_n}}{R}}{N}}}. \qquad \textbf{(34)}$$

Equation 34 represents the correlation between $\bar{\hat{\theta}}_n$ and θ_n after correcting for measurement error in $\bar{\hat{\theta}}_n$ and is, therefore, referred to as the "corrected correlation". Similarly, the variance difference between nominal and expected θ_n distributions was indexed by the "corrected variance ratio" :

$$\frac{S^2_{E[\hat{\theta}_n]}}{S_\theta{}^2} \approx \frac{S^2_{\bar{\theta}} - \dfrac{\sum_{n=1}^{N} \dfrac{s^2_{\hat{\theta}_n}}{R}}{N}}{S^2_\theta}. \qquad \textbf{(35)}$$

Finally, the corrected absolute mean difference between expected and nominal θ_n distributions was approximated by:

$$|\bar{X}_{E[\hat{\theta}_n]} - \bar{X}_{\theta}| \approx |\bar{X}_{\bar{\theta}} - \bar{X}_{\theta}|.$$

The corrected correlation, corrected variance ratio and corrected absolute mean difference were computed for the δ_i and τ_j parameters in an analogous fashion.

Results

Accuracy of θ_n Estimates. Figure 5 portrays the four accuracy measures associated with the θ_n estimates derived from the GUM. Each panel in Figure 5 presents the mean value of a different accuracy measure as a function of the number of subjects and number of items simulated in a given condition. All means were based on the 30 replications within a condition. The η^2 values shown at the top of each panel were calculated from a 6x5 between-subjects ANOVA in which the given accuracy measure served as the dependent variable and the number of items and the number of subjects served as the independent variables. These values indicate the proportion of the corrected total sum of squares that was attributed to the main effect of items (η^2_I), the main effect of subjects (η^2_S), or the interaction of items and subjects ($\eta^2_{I \times S}$).

Insert Figure 5 About Here

The average RMSE obtained under the different simulation conditions is shown in the upper left panel of Figure 5. The associated pattern of η^2 values indicated that the RMSE was almost totally determined by the number of items used to derive the θ_n estimates. The RMSE dropped dramatically as the number of items was increased from 5 to 10, and it dropped again when the number of items was increased to 15, albeit less substantially. There was little decrease in the

RMSE when the number of items was increased beyond 20, at which point, the average RMSE was equal to .313.

The upper right panel of Figure 5 shows the average correlation between nominal and estimated θ_n values obtained under alternative simulation conditions. These average correlations were greater than .959 in every condition, and thus, the estimated θ_n values were practically a linear function of nominal parameter values. Additionally, the small amount of variability that emerged in these average correlations was almost entirely a function of the number of items used to derive θ_n estimates. The size of the average correlation grew slightly as the number of items was increased to 20, at which point, the average correlation was equal to .992. Further increases in the number of items had little impact on the average correlation between estimated and nominal values.

The lower left panel Figure 5 portrays the average absolute mean difference between the estimated and nominal θ_n distributions for each condition. This measure was influenced by the number of items studied, the number of subjects sampled, and the interaction of these two factors. However, when 15 or more items were used to calculate θ_n estimates, then the average absolute mean difference between nominal and estimated distributions was always less than .03, regardless of the number of subjects studied. Moreover, this difference was generally insensitive to further increases in the number of items used.

The lower right panel of Figure 5 portrays the average variance ratio obtained in each condition. The fact that these average ratios were always greater than 1.0 indicated that the variance of the θ_n estimates was consistently larger than the variance of the nominal parameters. Moreover, the variance ratio was almost totally a function of the number of items used. The

variance ratio decreased as the number of items increased to 20, after which, further increases in the number of items led to only slight changes. The average variance ratio observed when 20 items were used to derive θ_n estimates was equal to 1.178.

Accuracy of δ_i Estimates. The accuracy measures for δ_i estimates are shown in Figure 6. The RMSE and variance ratio measures were both primarily determined by the number of items studied and these measures showed only small decreases when the number of items was increased beyond 20. The average value of the RMSE was equal to .217 when 20 items were studied, and the average variance ratio was equal to 1.164. The average correlation between nominal and estimated δ_i parameters was consistently near 1.0 in every condition studied.

Although the number of subjects had negligible impact on the accuracy of the δ_i estimates (relative to the impact of the number of items), the effects of subjects were in the expected direction. Specifically, the average RMSE and variance ratio generally decreased as the number of subjects increased.

Insert Figure 6 About Here

Accuracy of τ_j Estimates. Figure 7 portrays the accuracy measures for the τ_j estimates derived from the GUM. The RMSE, variance ratio and absolute mean difference measures were all primarily a function of the number of items studied. Moreover, the values of these accuracy measures consistently decreased as the number of items was increased to 20, after which further increases in the number of items had little impact. When 20 items were utilized, the average RMSE was equal to .143, the average absolute mean difference was equal to .111, and the average variance ratio was equal to 1.262. Again, the relatively small effect of sample size was in

the expected direction. The three accuracy measures generally improved as the sample size increased.

Insert Figure 7 About Here

The average correlation between estimated and nominal τ_j values was quite high in every condition (i.e., above .983), but the small amount of variance that did emerge in these average correlations was due to both the number of items studied and the number of subjects sampled. When the number of subjects was 1000 or more, then the average correlation increased slightly as the number of items approached 10, but further increases in the number of items had little effect on the size of these correlations. With smaller samples, the correlations continued to increase slightly as the number of items reached 20 to 25.

Bias in Parameter Estimates. The four measures of bias are plotted for the θ_n, δ_i, and τ_j estimates in Figures 8 through 10, respectively. The behavior of these four measures was very similar to that exhibited by the measures of accuracy, and therefore, only the most salient features will be noted here. The RMSB values indicated that the parameter estimates were generally biased, but that the degree of bias decreased as the test length increased. The reduction in bias was quite noticeable until the test reached 20 items in length, beyond which, further decreases were small but steady. With 20 calibrated items, the average RMSB was equal to .150, .201 and .129 for the θ_n, δ_i and τ_j parameters, respectively. There was also a very small main effect of sample size such that the RMSB decreased slightly as the sample grew to 500 subjects. Further increases in sample size led to negligible changes in RMSB.

Insert Figures 8 through 10 About Here

The corrected correlation was extremely high for all parameters in every condition studied. The correlation never fell below .988, and it was consistently above .999 when 15 or more items were calibrated. Therefore, the bias in parameter estimates was not due to a suboptimal correlation between expected values and nominal parameters.

The corrected variance ratio was consistently greater than 1.0 in every condition studied. This indicated that the unit of the expected parameter scale was inflated as compared to the nominal parameter scale. The degree of scale inflation was determined primarily by the number of items calibrated. The amount of inflation decreased moderately as the test length increased to 20 items, after which, subsequent decreases were relatively small. The average corrected variance ratio found with 20 items was equal to 1.154, 1.163 and 1.249 for the θ_n, δ_i and τ_j parameters, respectively. Like the RMSB, the corrected variance ratio was also slightly dependent on the number of subjects sampled, and the mean ratio decreased as the number of subjects grew to 500. Further increases in sample size led to negligible effects.

The corrected absolute mean difference for θ_n parameters was generally small, and it was consistently less than .016 whenever the number of items calibrated was 15 or greater. Therefore, differences in the locations of expected and nominal θ_n distributions contributed little to the bias depicted by the RMSB. In contrast, the corrected absolute mean difference for τ_j parameters was much larger and decreased steadily as the test length increased. Consequently, its behavior resembled that for the corrected variance ratio. This result was not surprising because scale inflation will typically be reflected by both the corrected absolute mean difference and the

corrected variance ratio associated with τ_j parameters.[5]

Accuracy of Standard Error Approximations. The model-based standard errors obtained from equations 23 and 24 are simple approximations and will generally be too small. The degree of underestimation inherent in these approximations was assessed by comparing them to the corresponding empirical standard errors (i.e., $s_{\hat{\theta}_n}$, $s_{\hat{\delta}_i}$ and $s_{\hat{\tau}_j}$) calculated from the simulation data. Model-based standard errors were derived by substituting nominal parameter values into equations 23 and 24. The ratio of model-based standard error to empirical standard error was computed for each parameter and then averaged over all parameters of a given type. The mean difference between model-based and empirical standard errors was also calculated for each set of parameters.

Insert Figure 11 About Here

Figure 11 portrays the results of the standard error comparisons. The panels on the left side of Figure 11 portray the mean ratio of standard errors for each parameter type as a function of the number of items and the number of subjects studied. The underestimation inherent in equations 23 and 24 was quite clear. The mean ratio was always less than 1 in every condition studied, and

---

[5] Recall that only those thresholds associated with the first C subjective response categories are estimated in the GUM. The remaining M-C thresholds are constrained by the symmetry assumption. Therefore, the C nominal τ_j parameters will not generally have a mean of 0. If the underlying continuum is inflated by some constant factor c, where $c > 1$, then the τ_j parameters will be multiplied by that same factor. Consequently, the mean of the τ_j distribution will increase by a factor of c and the variance will increase by a factor of c^2. In contrast, the GUM constrains the δ_i distribution to have a mean of 0, and therefore, inflation of the underlying continuum cannot affect the mean of the δ_i distribution. Similarly, when nominal θ_n values are sampled from a $N(0,\sigma)$ distribution, then scale inflation cannot change the expected value of that distribution.

it was often substantially less. However, the approximation generally improved as the number of test items grew larger, although this improvement was less consistent for τ_j parameters. With 20 test items, the average standard error ratio was equal to .894, .709 and .788 for the θ_n, δ_i, and τ_j parameters, respectively. The mean standard error ratio for θ_n estimates was also influenced by the number of subjects studied, and it typically grew larger as the number of subjects increased.

The panels on the right side of Figure 11 display the mean difference between model-based and empirical standard errors as a function of the number of items and subjects under study. Although the underestimation of standard errors was often substantial in a relative sense, it was generally small in a absolute sense. For example, when 20 items were studied, the average difference in standard errors was equal to -.029, -.028 and -.015 for the θ_n, δ_i, and τ_j parameters, respectively.

When taken together, these results suggest that estimates calculated from equations 23 and 24 will provide rough approximations to the true standard errors. Although other complex estimates might be more accurate (see Lord & Wingersky, 1985), these simple estimates exhibit little absolute error when tests of 20 or more items are administered to samples of 100 or more subjects. Such small inaccuracies are unlikely to make much difference in traditional attitude measurement situations.

Discussion

From a statistical perspective, the simulation results indicated that joint maximum likelihood (JML) estimates of GUM parameters will generally be biased. This finding was not surprising given the bias previously seen in JML estimates from other IRT models (Anderson, 1973; Jansen, van den Wollenberg & Wierda, 1988; Swaminathan & Gifford, 1983; van den Wollenberg, Wierda & Jansen, 1988; Wright & Douglas, 1977). In general, the degree of bias in GUM

estimates was inversely related to both test length and sample size, although the effects of sample size were very minor. This pattern of results suggested that JML estimates of GUM parameters may be asymptotically unbiased (i.e., consistent) as both the number of items and the number of subjects become large.

Interestingly, the bias in JML estimates manifested itself primarily in the unit of the resulting scale such that more bias led to an inflated unit of measurement. Such behavior implies that a bias correction factor could potentially be developed. Any correction factor would obviously be related to the number of items calibrated, but the role of other factors such as the number of response categories and the relative locations of items and examinees would also need to be evaluated.

From an applied perspective, the simulation results indicated that the JML algorithm can yield reasonably accurate estimates of GUM parameters with feasible data demands. The correlation between estimated and nominal parameters was always extremely high in every condition studied. Additionally, the inaccuracies observed in the JML estimates became quite small when the number of items studied approached 20, and this finding was generally independent of the number of subjects utilized. This does not mean that the number of subjects had no effect on parameter estimates. On the contrary, the accuracy of item parameter estimates typically increased as the number of subjects increased. However, the effects of sample size on accuracy were extremely small relative to the effects of test length, at least within the range of sample sizes studied here. Therefore, these results suggest that reasonably accurate estimates of GUM parameters can be obtained in many practical attitude measurement situations where the attitude questionnaire is based on as few as 15 to 20 6-category items and the number of subjects is 100 or more.

An Example With Real Data

Graded responses to Thurstone's (1932) attitude toward capital punishment scale were obtained from 245 University of South Carolina undergraduates. The 24 statements from the scale were presented randomly to each subject by means of a personal computer. Subjects responded to each statement using one of six response categories (i.e., "strongly disagree", "disagree", " slightly disagree", "slightly agree", "agree" and "strongly agree"). The empirical item characteristic curve associated with each attitude statement is given in Appendix B.

Dimensionality

Davison (1977) has shown that when graded responses follow a simple metric unfolding model, then the principal components of the interitem correlation matrix will suggest two primary dimensions and the component loadings will form a simplex pattern. Simulations of the GUM have suggested that the structure of the interitem correlation matrix will be similar when the model holds perfectly. (The component loadings will form a simplex-like structure, but the endpoints of the simplex will be folded inward.) Therefore, a principal components analysis of the interitem correlation matrix was conducted in an effort to identify those items that were least likely to conform to the unidimensionality assumption of the GUM. Conformability was operationally defined in terms of the item-level communality estimates derived from the first two principal components. Specifically, an item was discarded if less than 30 percent of its variation was determined from a two-component solution. This criterion, although somewhat arbitrary, appeared to be reasonable based on previous simulations. Seven items were discarded due to this requirement (i.e., statements 7, 11, 13, 17, 18, 20, and 22 from Appendix B). Responses to three of these items (statements 11, 13 and 20) appeared to be relatively independent of attitude toward

capital punishment. Responses to the other four statements were more or less consistent with the general hypothesis of an ideal point response process, but the variation in these responses was not adequately described by the first two principal components.

Eliminating the Most Ill-fitting Items from the Final Scale

After an initial calibration of model estimates was obtained, the most ill-fitting items were identified on the basis of two criteria. First, Wright and Masters' (1982) item fit *t* statistic (i.e., the infit statistic) was computed for each item. This statistic indexes the degree to which item-level responses fail to conform to the model. Wright and Masters claim that the item fit *t* statistic follows a standard t-distribution when applied to data that conform to a cumulative model. However, the distribution of the statistic under the GUM is not known. Therefore, this statistic was used as a heuristic to identify the most ill-fitting items. The second criterion used to identify deviant items was based on the degree to which an item's location on the attitude continuum appeared inconsistent with its content. Although this type of evaluation was obviously subjective, only those items with clearly questionable locations were removed. Together, these two heuristics led to the removal of 5 items (i.e.,statements 1, 5, 8, 15 and 16 from Appendix B). The characteristics of these items were consistent with the general hypothesis of an ideal point response process, yet their fit to the GUM (i.e., a specific model of that process) was questionable. The removal of these items left 12 items for the final calibration of model parameters.

Item Location Estimates

The item location estimates for the 12 selected items are shown in Table 1 along with their approximate standard errors. Items have been ordered on the basis of their location estimates,

and the sentiment expressed by each item is consistent with this same ordering. The consistency observed between item sentiment and item location provides an intuitive check on the adequacy of the model.

Insert Table 1 About Here

There was a pronounced gap between the estimated locations of statements 8 and 9. This was presumably due to a lack of items in the initial statement pool that reflected these intermediate opinions. Ordinarily, a test developer would attempt to calibrate a much larger set of items and then choose a subset of items from the final calibration that represented alternative regions of the latent continuum in a fairly uniform manner.

Attitude Estimates

The distribution of individual attitude estimates is shown in Figure 12. The mean attitude estimate was equal to 1.229 with a standard deviation of 1.204. The median attitude estimate of 1.286 fell between the statements "I do not believe in capital punishment, but it is not practically advisable to abolish it" and "We must have capital punishment for some crimes." Thus, these subjects begrudgingly supported capital punishment to at least some extent.

Insert Figure 12 About Here

Subjective Response Category Threshold Estimates

Figure 13 illustrates the subjective response category probability curves that were estimated with the GUM. In Figure 13, the locations of estimated subjective category thresholds are indicated by vertical lines. The thresholds were successively ordered and formed a series of

intervals on the latent continuum in which a particular subjective response was most likely. The intervals associated with a strongly agree response (from either below or above) were relatively wider than those for the remaining subjective responses. However, this result may have been due to the fact that a majority of individuals were located within the gap on the latent continuum bordered by items 8 and 9, and consequently, there was less information about the exact width of the strongly agree intervals.

Insert Figure 13 About Here

Global Model Fit

Although model fit statistics have not yet been developed for the GUM, a descriptive analysis was conducted to gain an intuitive understanding of the global model fit to the capital punishment data. In this analysis, the difference between each individual's attitude estimate and each estimated item location (i.e., $\hat{\theta}_n - \hat{\delta}_i$) was calculated. These differences were then sorted and divided into 70 homogeneous groups of equal size (n=42). The GUM expected values and the observed scores were both averaged across the individual-item pairs within each homogenous group. In this way, much of the random variation was averaged out of the observed scores. A strong linear relationship emerged between average observed scores and the average expected values as evidenced by a correlation of .993. Thus, the GUM fit these data reasonably well.

Insert Figure 14 About Here

Figure 14 arrays the average expected values and average observed scores as a function of the mean $\hat{\theta}_n - \hat{\delta}_i$ value within each homogenous group. The average observed scores are given by

squares whereas the average expected values are represented by the curve. As seen in the plot, the GUM arranged both individuals and items on the attitude continuum so that the average expected value and the average observed score both increased as the mean difference between individual and item locations approached zero. Furthermore, the fit of the GUM appeared reasonable throughout the range of attitudes exhibited in the sample.

Insert Figure 15 About Here

As an additional check on the adequacy of the model, the variance of observed scores within each homogenous group was compared to the corresponding variance predicted by the model. These variances are shown in Figure 15 as a function of the average $\hat{\theta}_n - \hat{\delta}_i$ value within each homogenous group. The observed variances are represented by squares and the model variances are given by the solid curve. As shown in Figure 15, the observed score variances generally mimicked the basic pattern exhibited by the model variances. This behavior led to a .715 correlation between observed score variances and model variances. Although the strength of this relationship was less than that found with expected values, it was still substantial, and it suggested that the GUM was a reasonable model for the capital punishment data.

Extensions and Future Research

Alternative Parameterizations of the GUM

Alternative versions of the GUM can be developed simply by reparameterizing the cumulative model which forms the basis for the unfolding mechanism. For example, if one presumes that subjective responses follow a partial credit model (Masters, 1982) rather than a rating scale

model, then a partial credit GUM (PC-GUM) can be derived as:

$$Pr[X_{ni} = x] = \frac{\exp[x(\theta_n - \delta_i) - \sum_{j=0}^{x} \tau_{ji}] + \exp[(M-x)(\theta_n - \delta_i) - \sum_{j=0}^{x} \tau_{ji}]}{\sum_{w=0}^{C} [\exp[w(\theta_n - \delta_i) - \sum_{j=0}^{w} \tau_{ji}] + \exp[(M-w)(\theta_n - \delta_i) - \sum_{j=0}^{w} \tau_{ji}]]}, \quad (36)$$

where τ_{ji} are subjective response category thresholds that are allowed to vary across items and the remaining terms are defined as in equation 4. This model is appropriate if the scale associated with subjective responses varies across items.

An alternative GUM can be developed by constraining threshold parameters from an underlying rating scale model. Specifically, if one assumes that successive subjective response category thresholds are equally distant from each other, then subjective responses can be modeled with Andrich's (1982) "constant unit" version of the rating scale model. This would give rise to a constant unit GUM (CU-GUM):

$$Pr[X_{ni} = x] = \frac{\exp[x(M-x)\lambda + x(\theta_n - \delta_i)] + \exp[x(M-x)\lambda + (M-x)(\theta_n - \delta_i)]}{\sum_{w=0}^{C} [\exp[w(M-w)\lambda + w(\theta_n - \delta_i)] + \exp[w(M-w)\lambda + (M-w)(\theta_n - \delta_i)]]}, \quad (37)$$

where λ is a "unit" parameter that is equal to half the distance between successive thresholds and the remaining terms are defined as in equation 4. At a conceptual level, this model is appropriate if subjective responses to attitude items are on an equal interval scale. A similar assumption is that subjective responses are on an equal interval scale, but the scale unit changes with each item. In this case, the subjective responses could be modeled with Andrich's (1982) "multiple unit" version of the rating scale model. This would yield a multiple unit GUM (MU-GUM):

$$Pr[X_{ni} = x] = \frac{\exp[x(M-x)\lambda_i + x(\theta_n - \delta_i)] + \exp[x(M-x)\lambda_i + (M-x)(\theta_n - \delta_i)]}{\sum_{w=0}^{C} [\exp[w(M-w)\lambda_i + w(\theta_n - \delta_i)] + \exp[w(M-w)\lambda_i + (M-w)(\theta_n - \delta_i)]]}, \quad \textbf{(38)}$$

where λ_i is a unit parameter that is allowed to vary across items.

Another useful GUM can be derived from Muraki's (1992) generalized rating scale model. The resulting model, referred to as the generalized GUM (G-GUM), includes a discrimination parameter that varies across items:

$$Pr[X_{ni} = x] = \frac{\exp\left[\alpha_i\left(x(\theta_n - \delta_i) - \sum_{j=0}^{x}\tau_j\right)\right] + \exp\left[\alpha_i\left((M-x)(\theta_n - \delta_i) - \sum_{j=0}^{x}\tau_j\right)\right]}{\sum_{w=0}^{C}\left(\exp\left[\alpha_i\left(w(\theta_n - \delta_i) - \sum_{j=0}^{w}\tau_j\right)\right] + \exp\left[\alpha_i\left((M-w)(\theta_n - \delta_i) - \sum_{j=0}^{w}\tau_j\right)\right]\right)}, \quad \textbf{(39)}$$

where α_i is the discrimination parameter for the *ith* item. This model is interesting because it allows for item-level trace lines that vary in their peakedness and dispersion while maintaining a constant set of subjective response category thresholds across items. In this regard, the G-GUM is conceptually similar to Thurstone's successive intervals procedure (Safir, 1937) for scaling attitude items in which the standard deviations of item scale value distributions are allowed to differ, but the response category boundaries remain fixed. However, the two models differ in that the successive intervals procedure posits a cumulative response mechanism whereas the G-GUM is based on an unfolding model.

Roberts (1995) investigated the ability to recover parameters from the CU-GUM and MU-

GUM using a joint maximum likelihood procedure analogous to that described in this paper. His results paralleled those reported here for the GUM. Specifically, he found that reasonably accurate estimates of all model parameters could be obtained with as few as 15 to 20 6-category items and as little as 100 subjects. An investigation of parameter recovery in the PC-GUM is currently underway, and preliminary results suggest that relatively large samples (e.g., 1000 subjects or more) are required to produce accurate measures of item-level thresholds. An investigation of parameter estimation for the G-GUM is planned for the near future.

Alternative Methods of Parameter Estimation

As noted above, the joint maximum likelihood procedure can yield reasonably accurate estimates of GUM parameters with minimal data demands. However, the method is computationally intensive, and the statistical consistency of the resulting estimates is generally questionable. Consequently, we have recently begun to explore estimation of GUM parameters with a marginal maximum likelihood procedure (Bock & Aitkin, 1981; Bock and Lieberman, 1970; Muraki, 1992). Preliminary results from this work appear promising and suggest that marginal maximum likelihood estimates will be more computationally efficient and more statistically sound than their joint maximum likelihood counterparts.

Summary

The GUM provides a means to simultaneously locate both items and persons on a unidimensional latent attitude continuum. It does so using an unfolding mechanism, and thus, it is consistent with Thurstone's (1928, 1931) view of the response process in attitude measurement. Moreover, it can be used in situations where responses are binary or graded, so

there is no need to dichotomize one's data, and lose potentially useful measurement information, in order to use the procedure. GUM parameters can be estimated in a reasonably accurate manner with relatively small data demands, and thus, the model should be applicable to a variety of practical attitude measurement situations. Lastly, alternative versions of the GUM can be easily adapted from a diverse set of cumulative IRT models which adds to the versatility of the approach.

References

Anderson, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31-44.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43(4)*, 561-573.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47(1)*, 105-113.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, *12(1)*, 33-51.

Andrich, D. (in press). A general hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17(3)*, 253-276.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (Part 5, pp. 397-497). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37(1)*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35(2)*, 179-197.

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12(1)*, 83-97.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons.

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika, 42*, 523-548.

Donoghue, J. R. (1994). An examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 41(4)*, 295-311.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. Suchman, P. E. Lazarsfeld, S. A. Star, & J. A. Gardner (Eds.), *Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55(4)*, 641-656.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15(2)*, 153-169.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6(3)*, 249-260.

Jansen, P. G. W., van den Wollenberg, A. L., & Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12(3)*, 297-306.

Kim, S., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59(3)*, 405-421.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, No. 140, 5-53.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 69-88). Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47(2)*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Roberts, J. S. (1995). *Item response theory approaches to attitude measurement*. Unpublished doctoral dissertation, University of South Carolina.

Roberts, J. S., Laughlin, J. E., & Wedell, D.H. (1996) *The Thurstone and Likert approaches to attitude measurement: Two alternative perspectives of the item response process*. Manuscript in preparation.

Safir, M. A. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika, 2(3)*, 179-198.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14(3)*, 299-311.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model (pp. 13-30). In D. Weiss (Ed.), *New horizons in testing*. New York, NY: Academic Press.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *The Journal of Educational Measurement, 13(3)*, 201-214.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33(4)*, 529-554.

Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Thurstone, L. L. (1932). *Motion pictures and attitudes of children*. Chicago, IL: University of Chicago Press.

van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement, 12(3)*, 307-313.

van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists.* Amsterdam: CT Press.

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model can be used instead. *Applied Psychological Measurement, 18(2)*, 97-110.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281-294.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52(2)*, 275-291.

Acknowledgments

This report was based partially on research from the first author's doctoral dissertation which was conducted at the University of South Carolina under the direction of the second author. We are grateful to Thomas Cafferty, Huynh Huynh and Douglas Wedell for their critical input during the dissertation process. We would also like to thank John Donoghue and Robert Mislevy for their helpful comments on earlier drafts of this report. Lastly, we have only recently learned that David Andrich has independently proposed an unfolding model for graded responses that is very similar to the GUM (Andrich, in press). In contrast to our work, however, Andrich's proposal does not address parameter estimation nor does it demonstrate the utility of his model with a real data application.

Author's Address

Correspondence should be directed to James S. Roberts, Educational Testing Service, Rosedale Road, Mail Stop 02-T, Princeton, NJ 08541. Electronic mail may be sent via Internet to JROBERTS@ETS.ORG.

Table 1

Scale Values for 12 Capital Punishment Statements from the Final Calibration

Statement	$\hat{\delta}_i$	$\hat{\sigma}_{\hat{\delta}_i}$
1) I do not believe in capital punishment under any circumstances.	-1.859	.083
2) Capital punishment is absolutely never justified.	-1.434	.071
3) Capital punishment is not necessary in modern civilization.	-1.378	.069
4) We can't call ourselves civilized as long as we have capital punishment.	-1.298	.068
5) Execution of criminals is a disgrace to civilized society.	-1.183	.065
6) Life imprisonment is more effective than capital punishment.	-1.081	.064
7) I don't believe in capital punishment but I'm not sure it isn't necessary.	-0.828	.061
8) I do not believe in capital punishment but it is not practically advisable to abolish it.	-0.604	.059
9) We must have capital punishment for some crimes.	2.049	.064
10) Capital punishment is just and necessary.	2.448	.061
11) Capital punishment gives the criminal what he deserves.	2.463	.061
12) Capital punishment should be used more often than it is.	2.704	.061

Figure Captions

Figure 1. Subjective response category probability curves for a hypothetical item under the GUM. The item has 4 observable response categories which leads to 8 subjective response categories. There is a probability curve corresponding to each of the subjective response categories. Subjective response category thresholds are located at -1.3, -.7, -.3, 0, .3, .7 and 1.3.

Figure 2. Observed response category probability curves for a hypothetical item under the GUM. There is a probability curve corresponding to each of the 4 observable response categories.

Figure 3. The expected value of an observable response to a hypothetical item under the GUM. The item has 4 observable response categories which are scored from 0 to 3 where larger scores are indicative of higher levels of agreement.

Figure 4. The standard error of a GUM attitude estimate (i.e., a θ_n estimate) based on a single individual's response to a single hypothetical item with 6 response categories. Each curve corresponds to a different value of ψ, where ψ is the distance between successive (equally spaced) subjective response category thresholds.

Figure 5. Mean accuracy measures associated with θ_n estimates from the GUM. Each mean is based on 30 replications within a given condition defined by the number of items and the number of subjects simulated. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 6. Mean accuracy measures associated with δ_i estimates from the GUM. Each mean is based on 30 replications within a given condition defined by the number of items and the number

of subjects simulated. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 7. Mean accuracy measures associated with τ_j estimates from the GUM. Each mean is based on 30 replications within a given condition defined by the number of items and the number of subjects simulated. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 8. Bias measures associated with θ_n estimates from the GUM. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 9. Bias measures associated with δ_i estimates from the GUM. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 10. Bias measures associated with τ_j estimates from the GUM. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 11. Comparisons between model-based estimates and empirical estimates of standard errors for GUM parameters. Comparisons have been averaged over all parameters of a given type. Each line, along with the associated symbol, corresponds to a given level of subjects where ✳=100 subjects, ○=200 subjects, □=500 subjects, △=1000 subjects and ●=2000 subjects.

Figure 12. Distribution of GUM attitude estimates (i.e., θ_n estimates) derived from responses to the 12 capital punishment items used in the final calibration.

Figure 13. Subjective response category probability curves derived from responses to the 12 capital punishment items used in the final calibration. The vertical lines indicate the locations of the subjective response category thresholds.

Figure 14. The relationship between average observed item responses and average expected item responses under the GUM portrayed as a function of $\theta_n - \delta_i$.

Figure 15. The relationship between observed score variance and variance predicted by the GUM as a function of $\theta_n - \delta_i$.

Figure 1

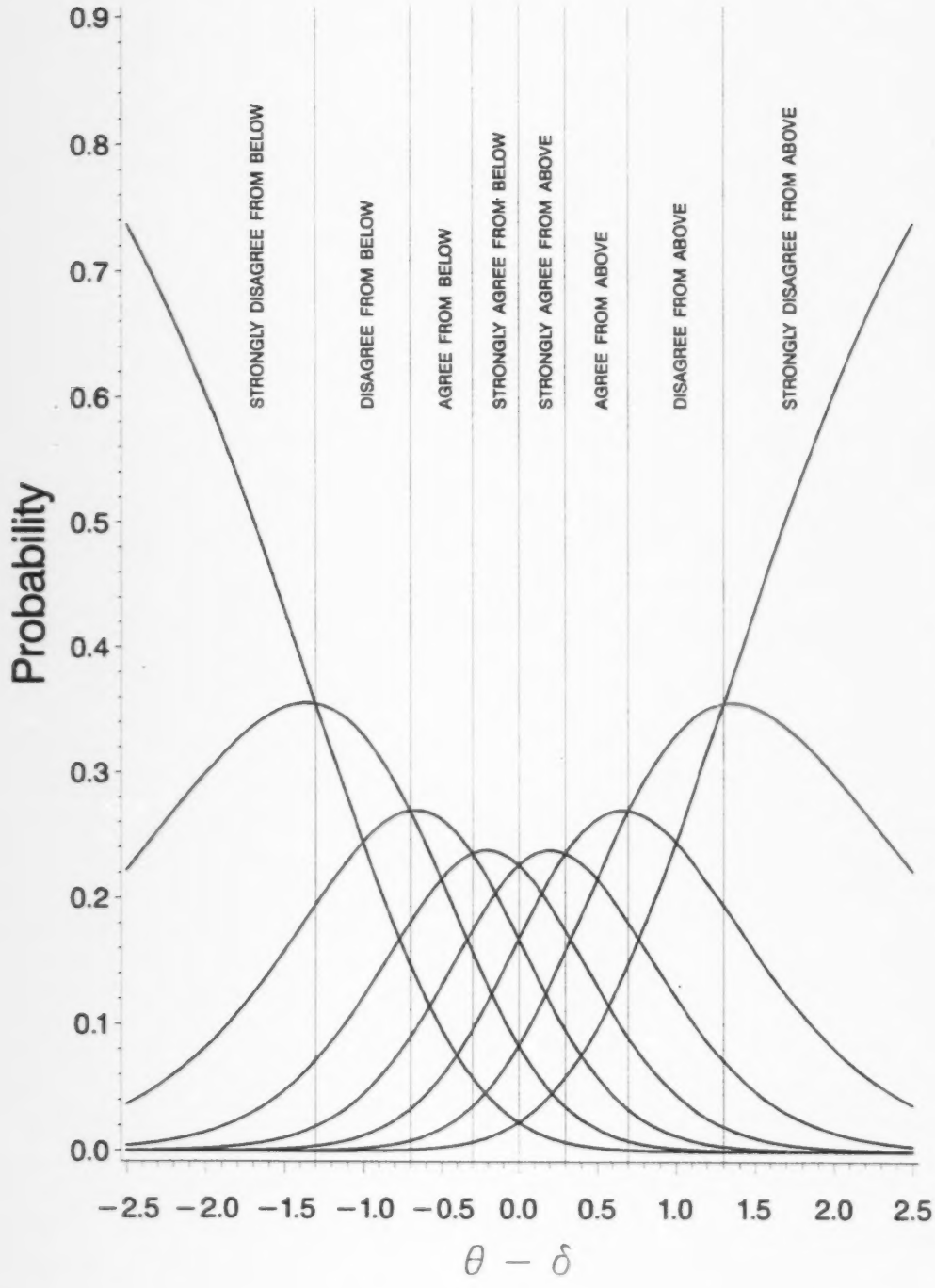

Probability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
STRONGLY DISAGREE FROM BELOW
DISAGREE FROM BELOW
AGREE FROM BELOW
STRONGLY AGREE FROM BELOW
STRONGLY AGREE FROM ABOVE
AGREE FROM ABOVE
DISAGREE FROM ABOVE
STRONGLY DISAGREE FROM ABOVE
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
θ − δ

Figure 2

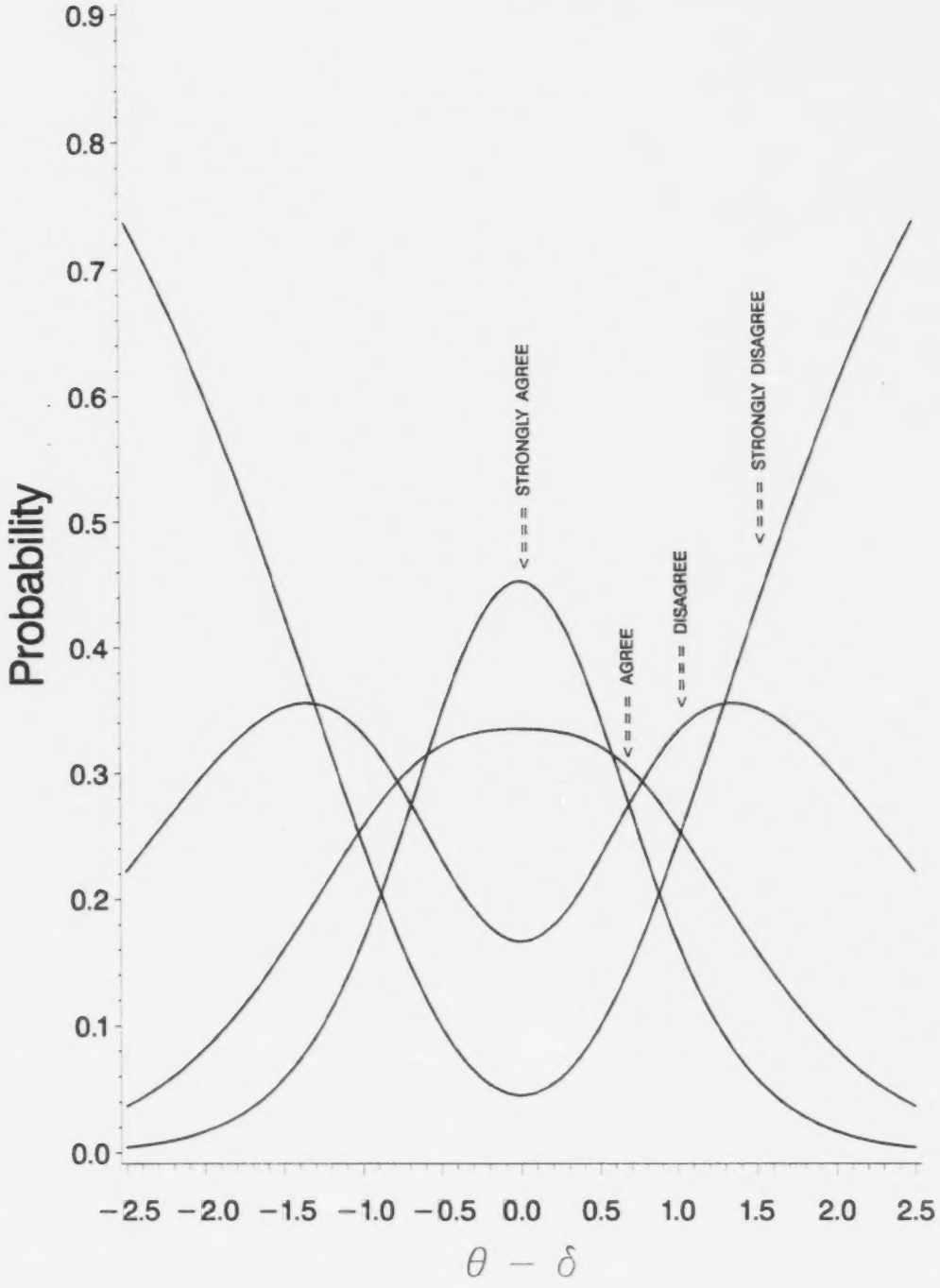

Probability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
<= = = STRONGLY AGREE
<= = = AGREE
<= = = DISAGREE
<= = = STRONGLY DISAGREE
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
θ − δ

Figure 3

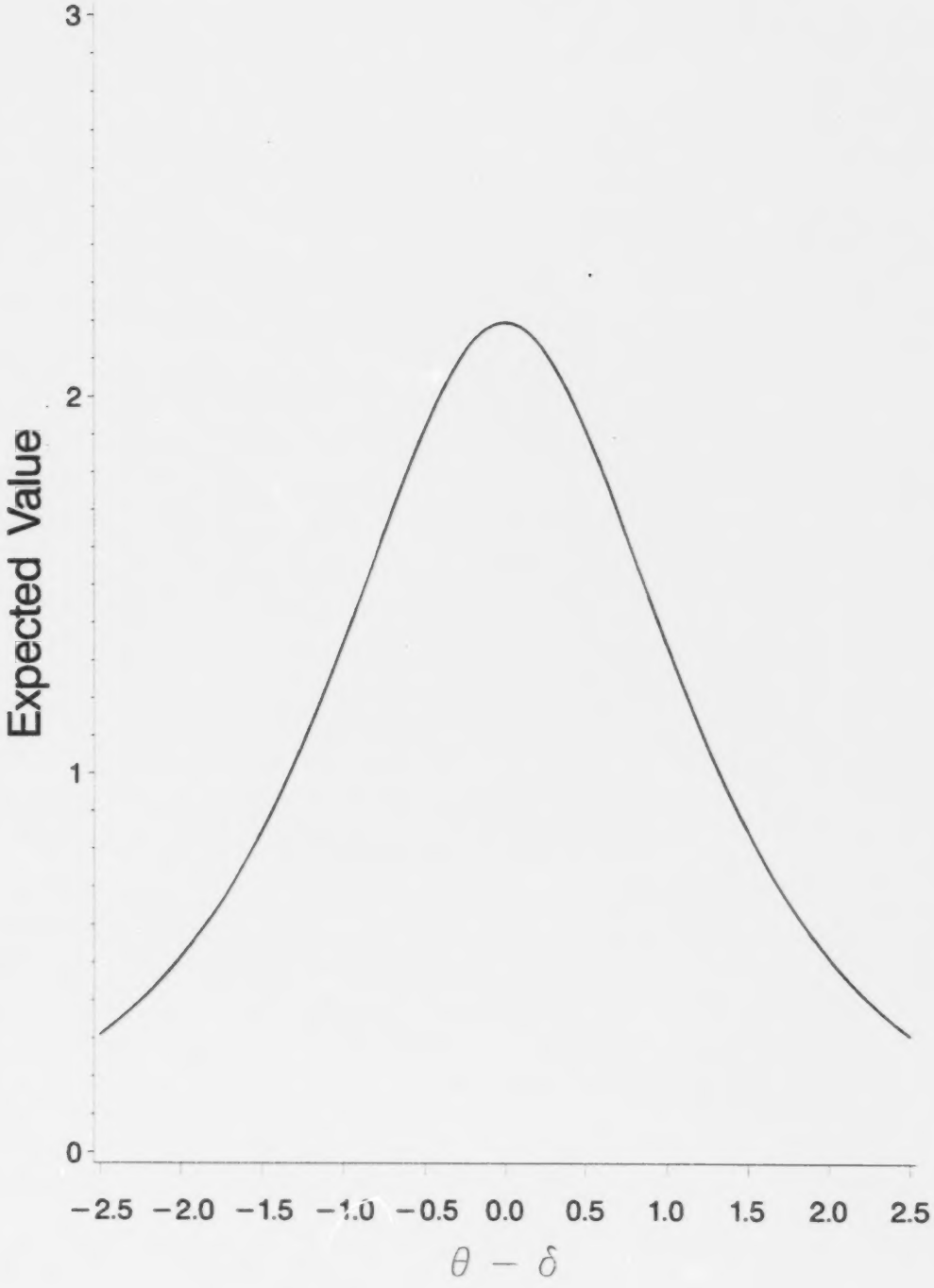
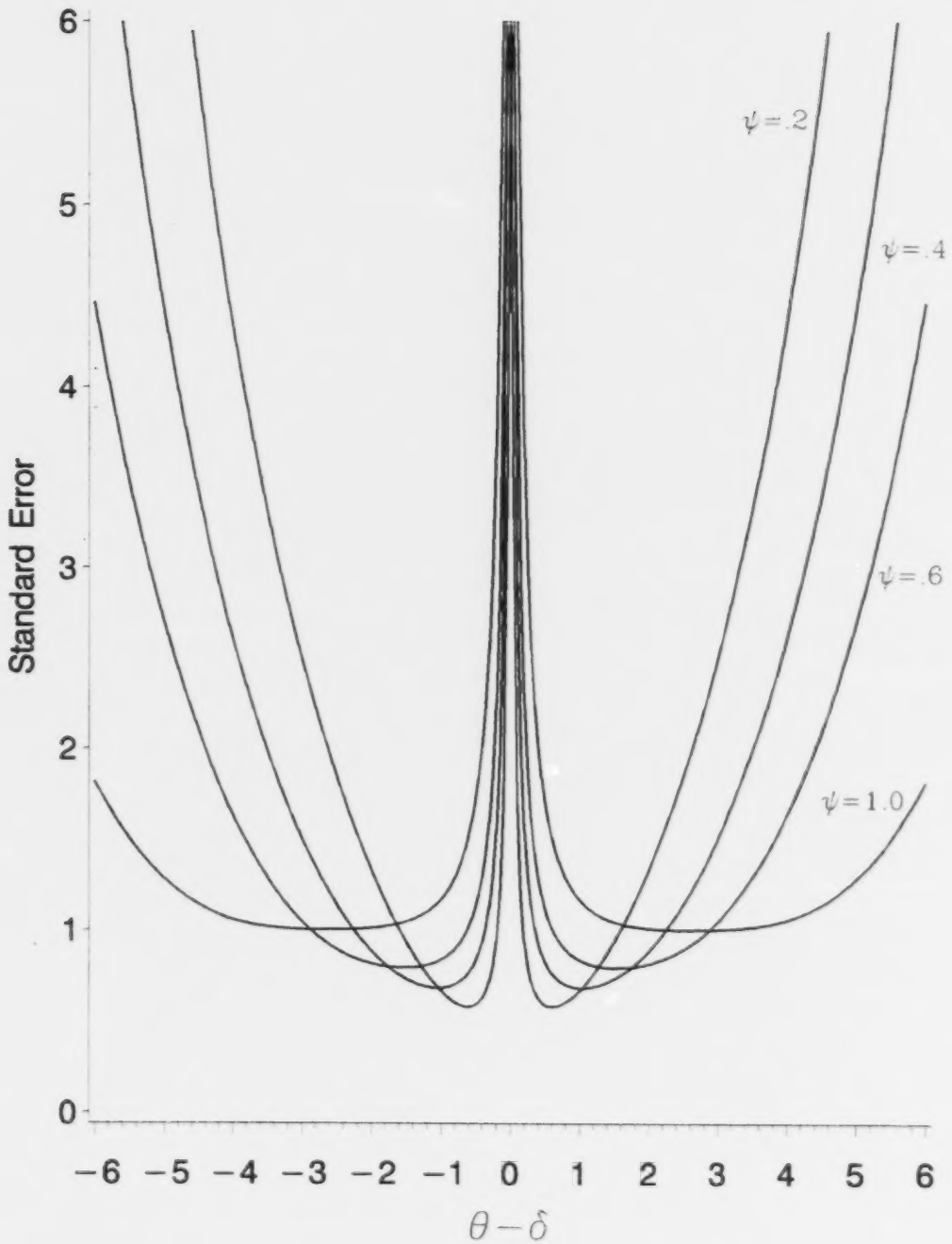
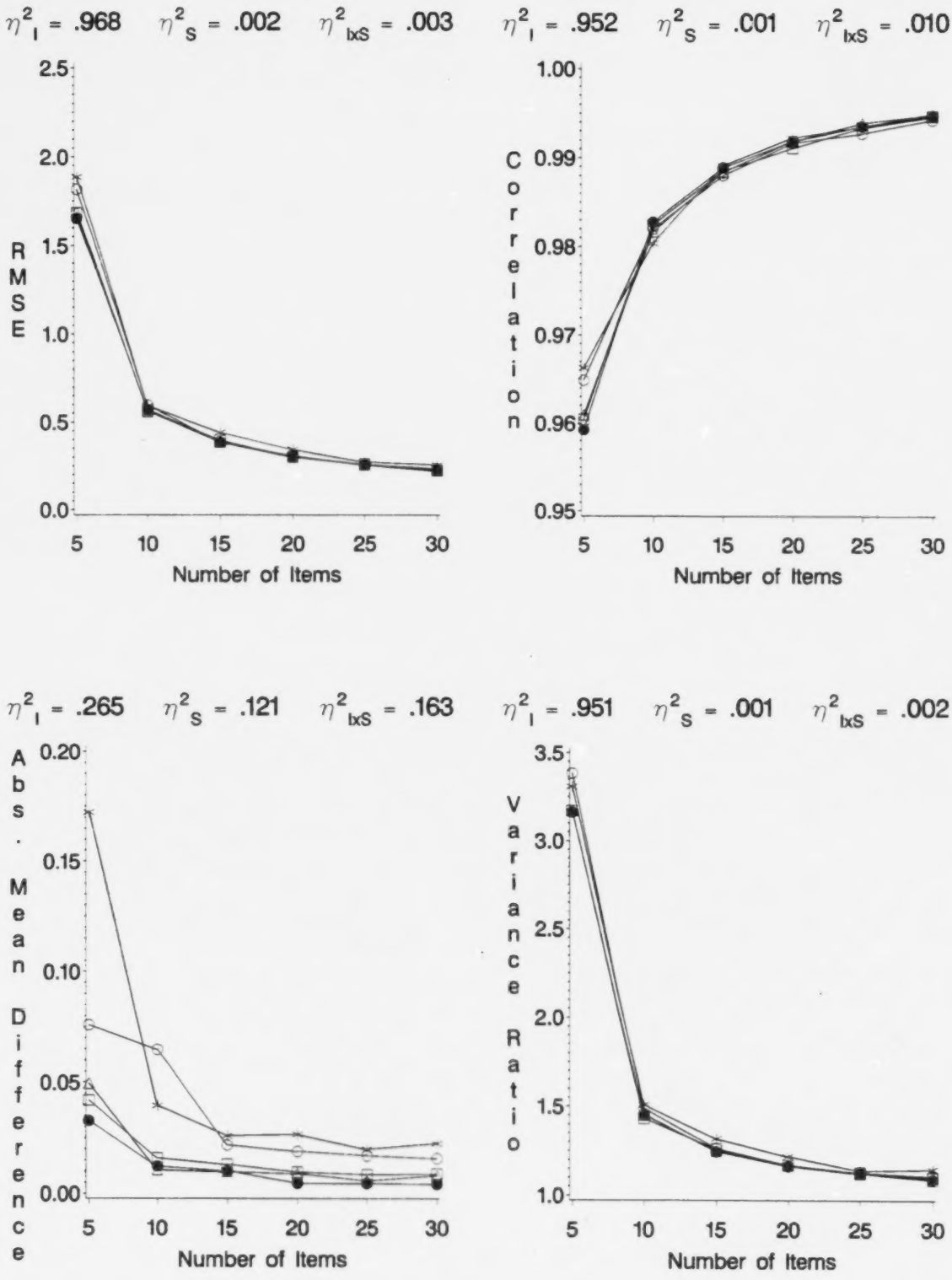
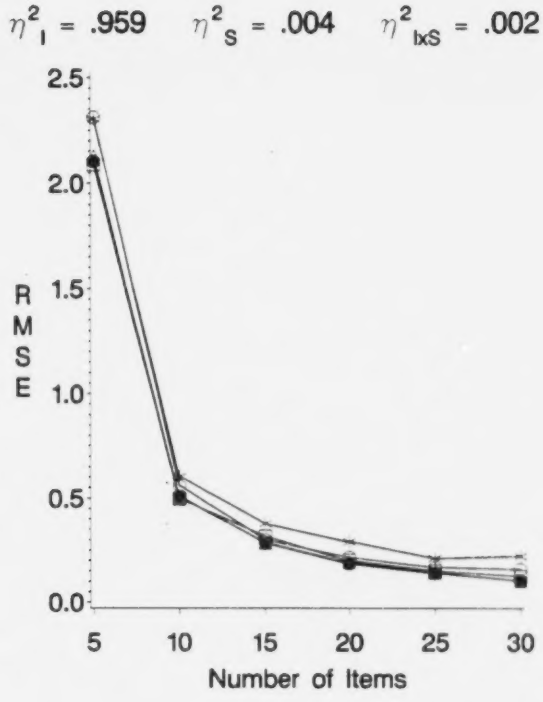
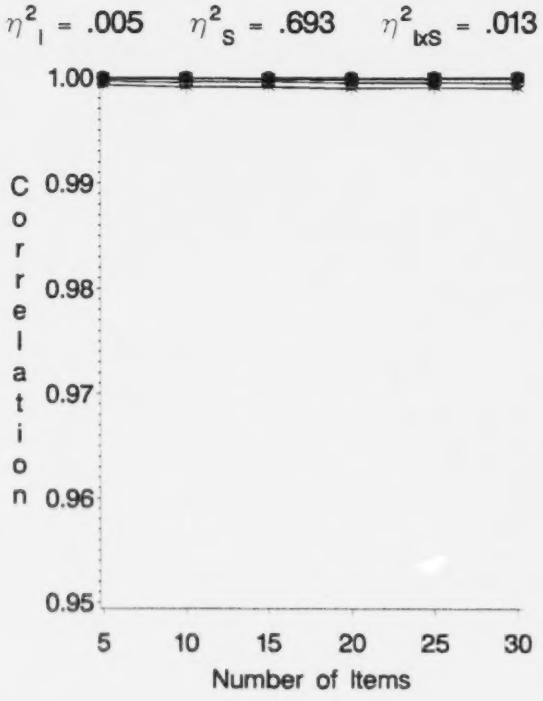


Figure 4


Standard Error
6
5
4
3
2
1
0
$\psi = .2$
$\psi = .4$
$\psi = .6$
$\psi = 1.0$
−6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6
$\theta - \delta$


η^2_I = .968 η^2_S = .002 η^2_{IxS} = .003
RMSE
2.5
2.0
1.5
1.0
0.5
0.0
5 10 15 20 25 30
Number of Items
η^2_I = .952 η^2_S = .001 η^2_{IxS} = .010
Correlation
1.00
0.99
0.98
0.97
0.96
0.95
5 10 15 20 25 30
Number of Items
η^2_I = .265 η^2_S = .121 η^2_{IxS} = .163
Abs. Mean Difference
0.20
0.15
0.10
0.05
0.00
5 10 15 20 25 30
Number of Items
η^2_I = .951 η^2_S = .001 η^2_{IxS} = .002
Variance Ratio
3.5
3.0
2.5
2.0
1.5
1.0
5 10 15 20 25 30
Number of Items





There were no absolute mean difference measures because both nominal and and estimated item locations were centered at zero.

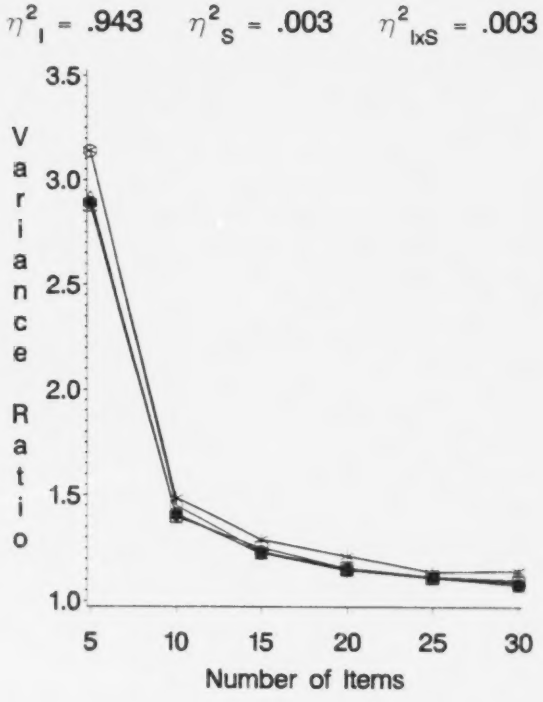


Figure 7

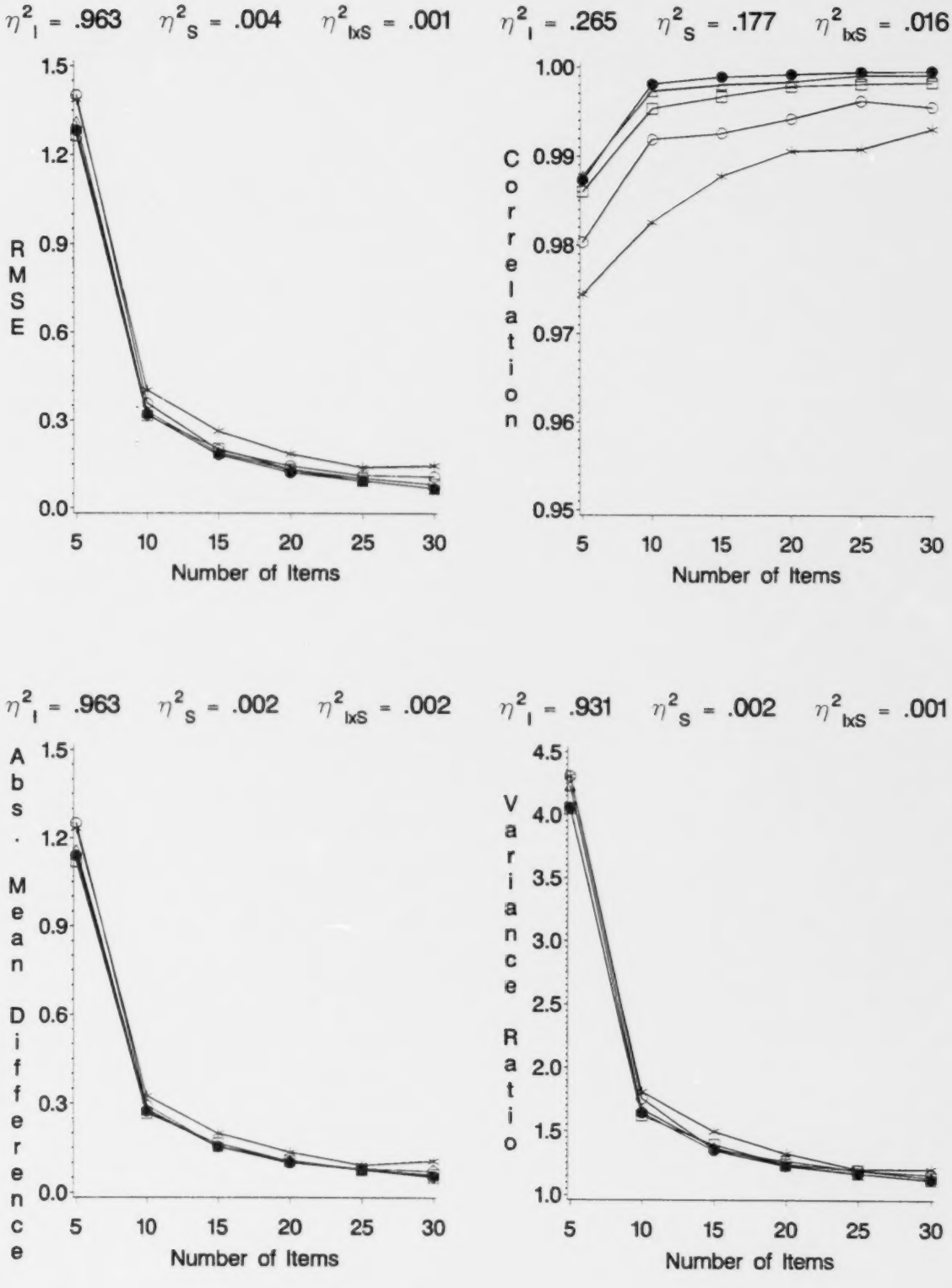
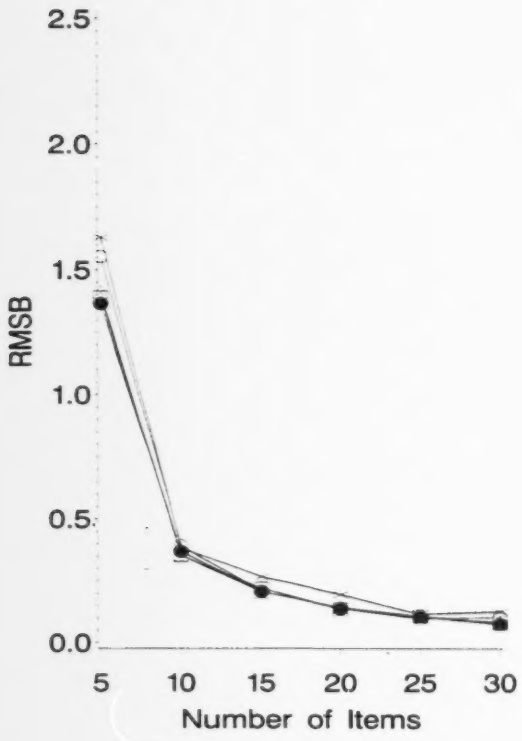
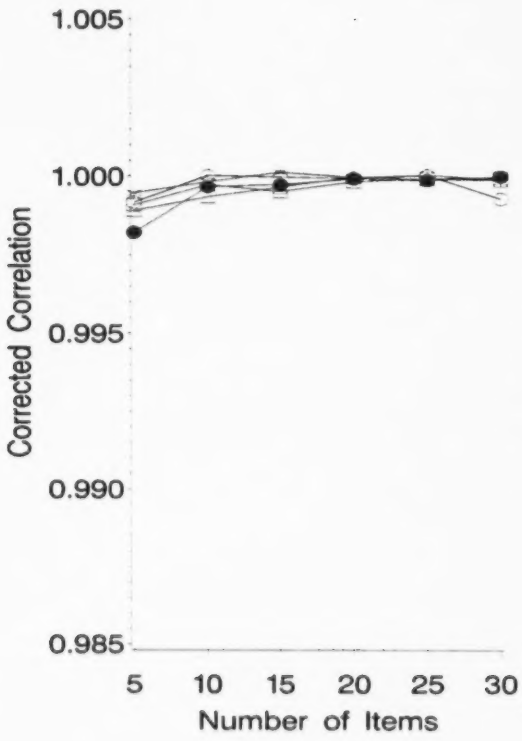
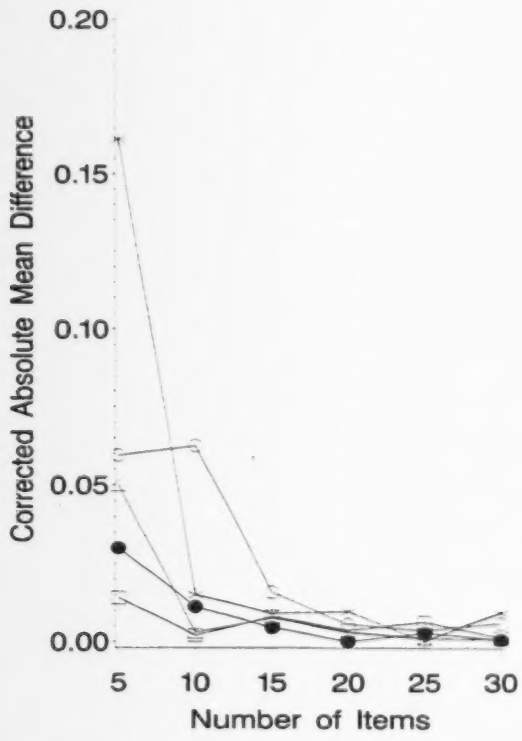
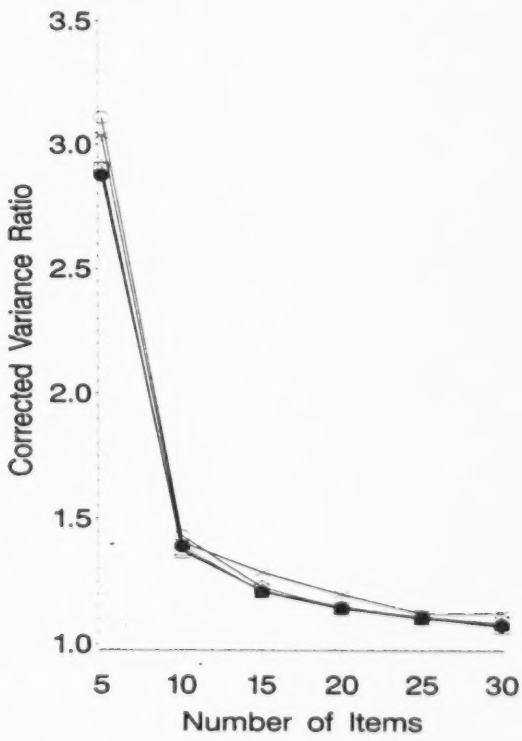

η^2_I = .963 η^2_S = .004 η^2_{IxS} = .001
RMSE
1.5
1.2
0.9
0.6
0.3
0.0
5 10 15 20 25 30
Number of Items
η^2_I = .265 η^2_S = .177 η^2_{IxS} = .016
Correlation
1.00
0.99
0.98
0.97
0.96
0.95
5 10 15 20 25 30
Number of Items
η^2_I = .963 η^2_S = .002 η^2_{IxS} = .002
Abs. Mean Difference
1.5
1.2
0.9
0.6
0.3
0.0
5 10 15 20 25 30
Number of Items
η^2_I = .931 η^2_S = .002 η^2_{IxS} = .001
Variance Ratio
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
5 10 15 20 25 30
Number of Items

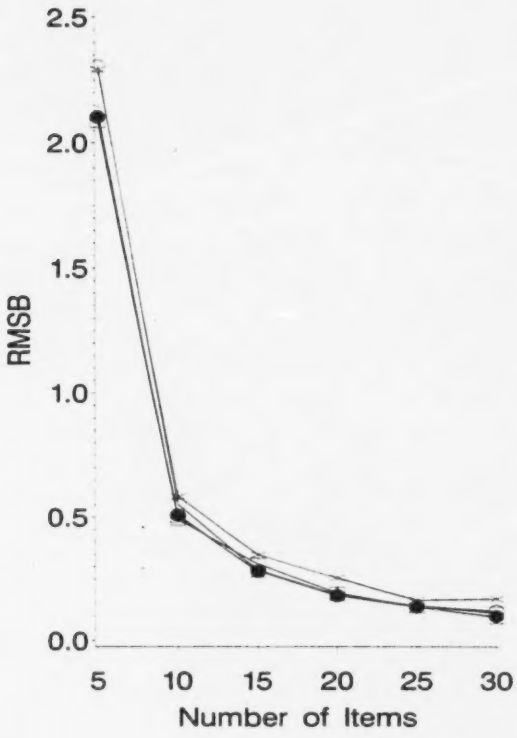
Figure 8

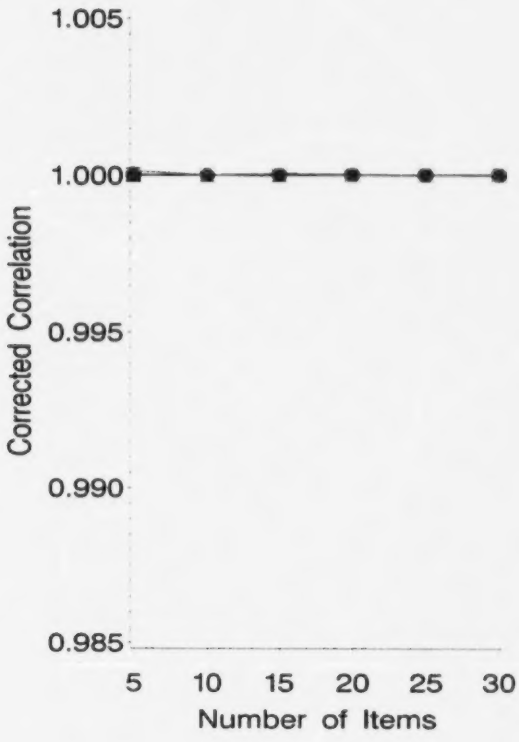

RMSB
2.5
2.0
1.5
1.0
0.5
0.0
5
10
15
20
25
30
Number of Items


Corrected Correlation
1.005
1.000
0.995
0.990
0.985
5
10
15
20
25
30
Number of Items


Corrected Absolute Mean Difference
0.20
0.15
0.10
0.05
0.00
5
10
15
20
25
30
Number of Items


Corrected Variance Ratio
3.5
3.0
2.5
2.0
1.5
1.0
5
10
15
20
25
30
Number of Items





There were no corrected absolute mean difference measures because both nominal and expected item locations were centered at zero.

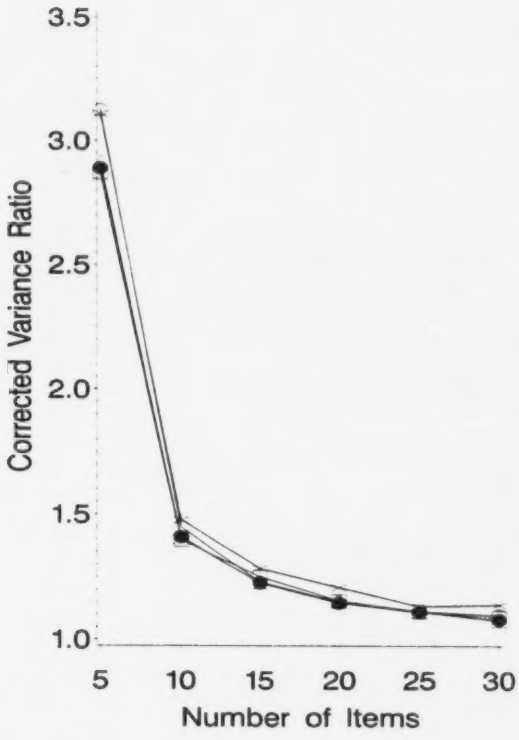
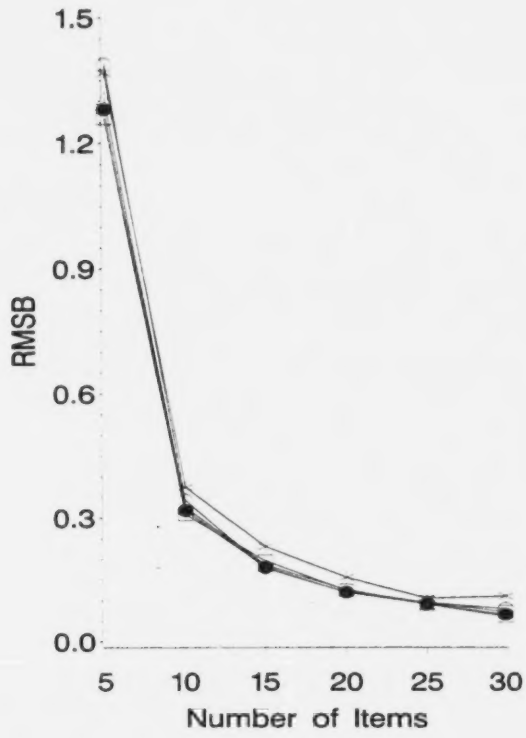
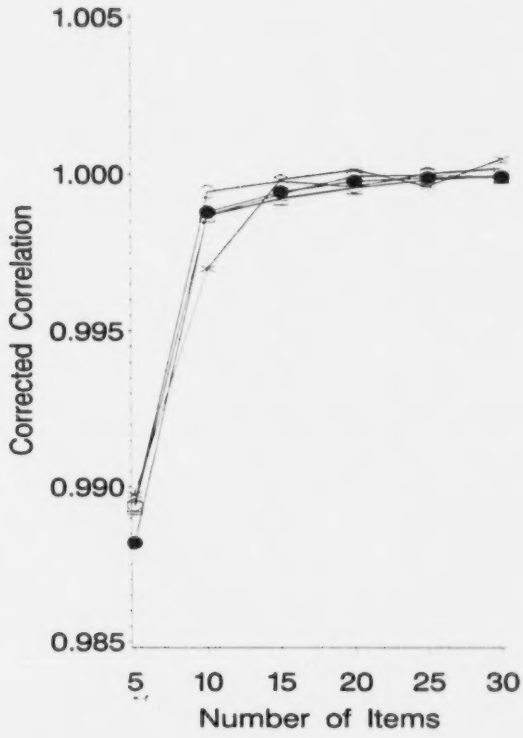
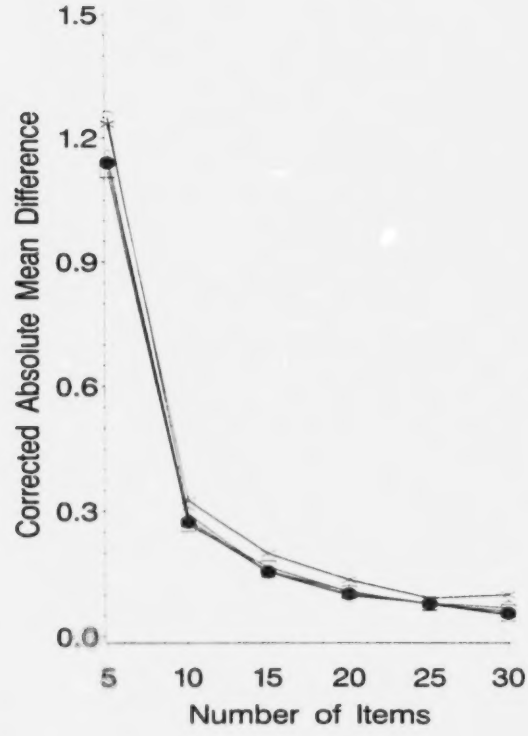


Figure 10


RMSB
1.5
1.2
0.9
0.6
0.3
0.0
5
10
15
20
25
30
Number of Items


Corrected Correlation
1.005
1.000
0.995
0.990
0.985
5
10
15
20
25
30
Number of Items


Corrected Absolute Mean Difference
1.5
1.2
0.9
0.6
0.3
0.0
5
10
15
20
25
30
Number of Items

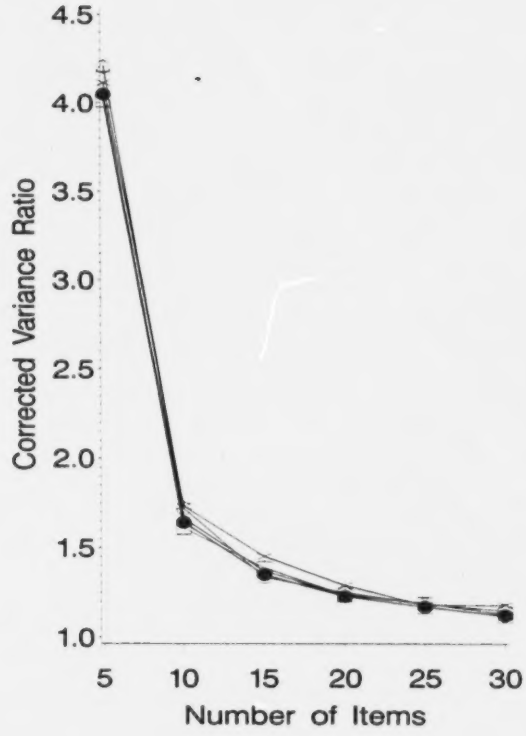

Corrected Variance Ratio
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
5
10
15
20
25
30
Number of Items

Figure 11

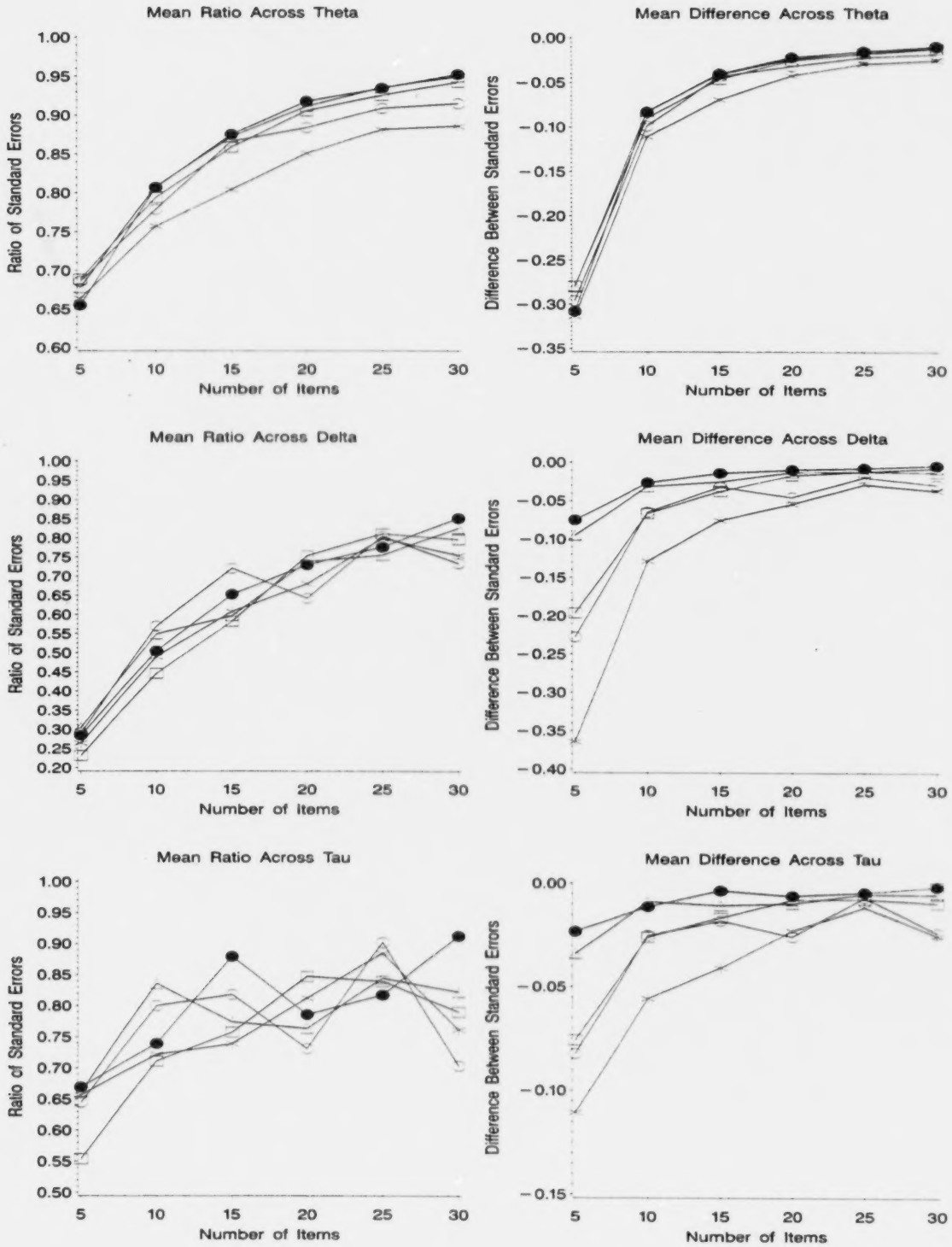

Mean Ratio Across Theta
Ratio of Standard Errors
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
15
20
25
30
Number of Items
Mean Difference Across Theta
Difference Between Standard Errors
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
Mean Ratio Across Delta
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Mean Difference Across Delta
−0.40
Mean Ratio Across Tau
Mean Difference Across Tau

Figure 12

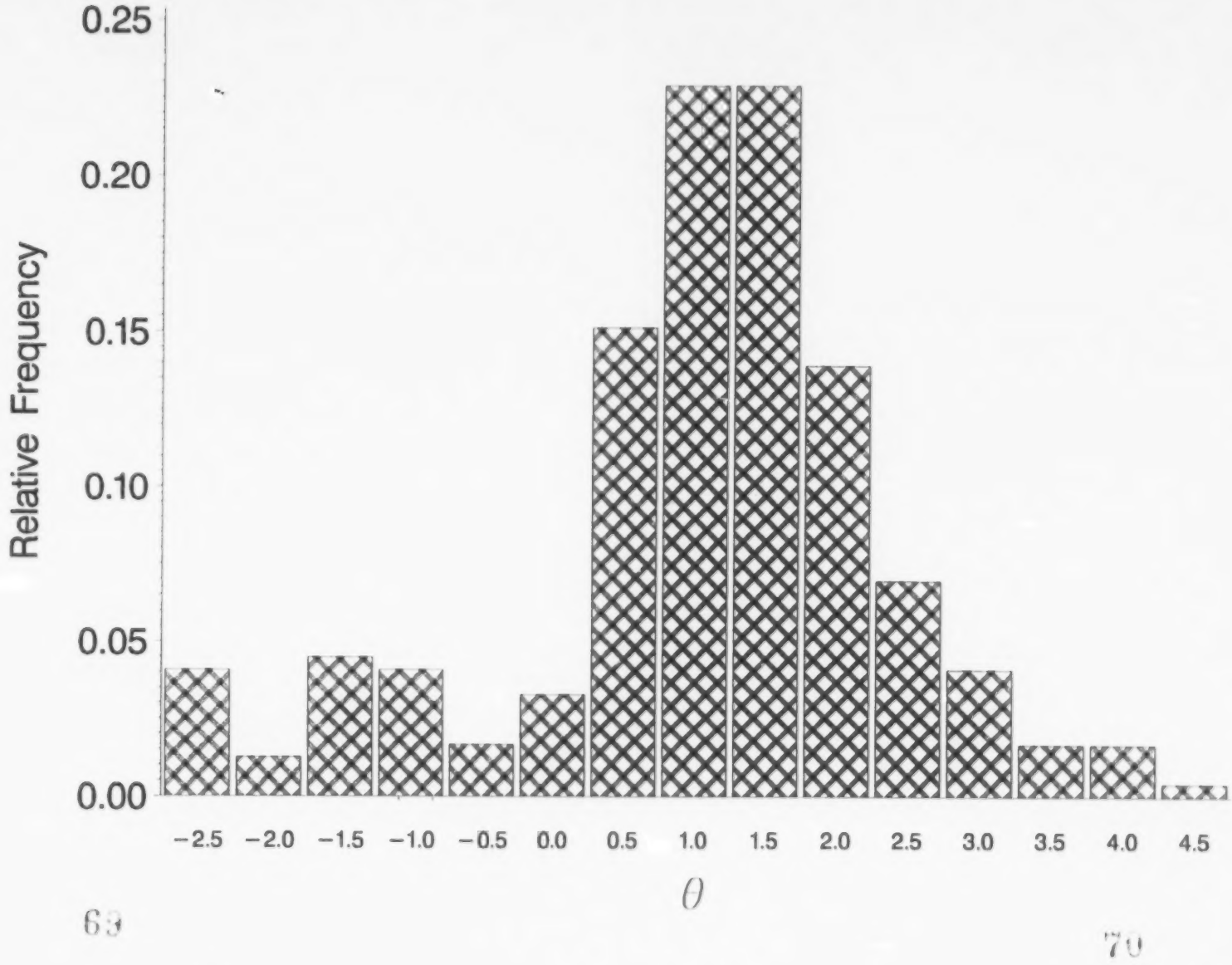

Relative Frequency
0.25
0.20
0.15
0.10
0.05
0.00
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
θ

Figure 13

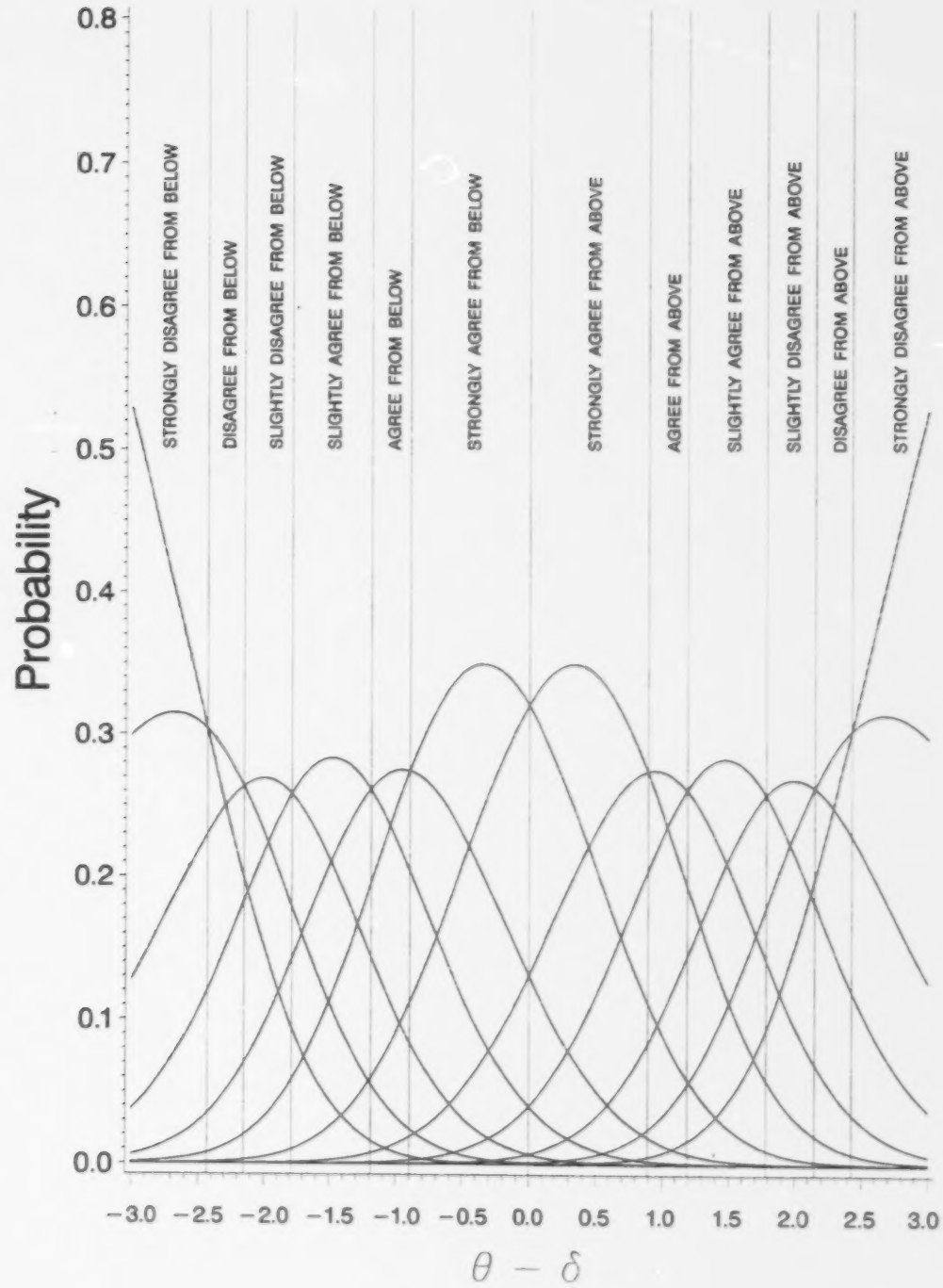

Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
STRONGLY DISAGREE FROM BELOW
DISAGREE FROM BELOW
SLIGHTLY DISAGREE FROM BELOW
SLIGHTLY AGREE FROM BELOW
AGREE FROM BELOW
STRONGLY AGREE FROM BELOW
STRONGLY AGREE FROM ABOVE
AGREE FROM ABOVE
SLIGHTLY AGREE FROM ABOVE
SLIGHTLY DISAGREE FROM ABOVE
DISAGREE FROM ABOVE
STRONGLY DISAGREE FROM ABOVE
−3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
θ − δ

Figure 14

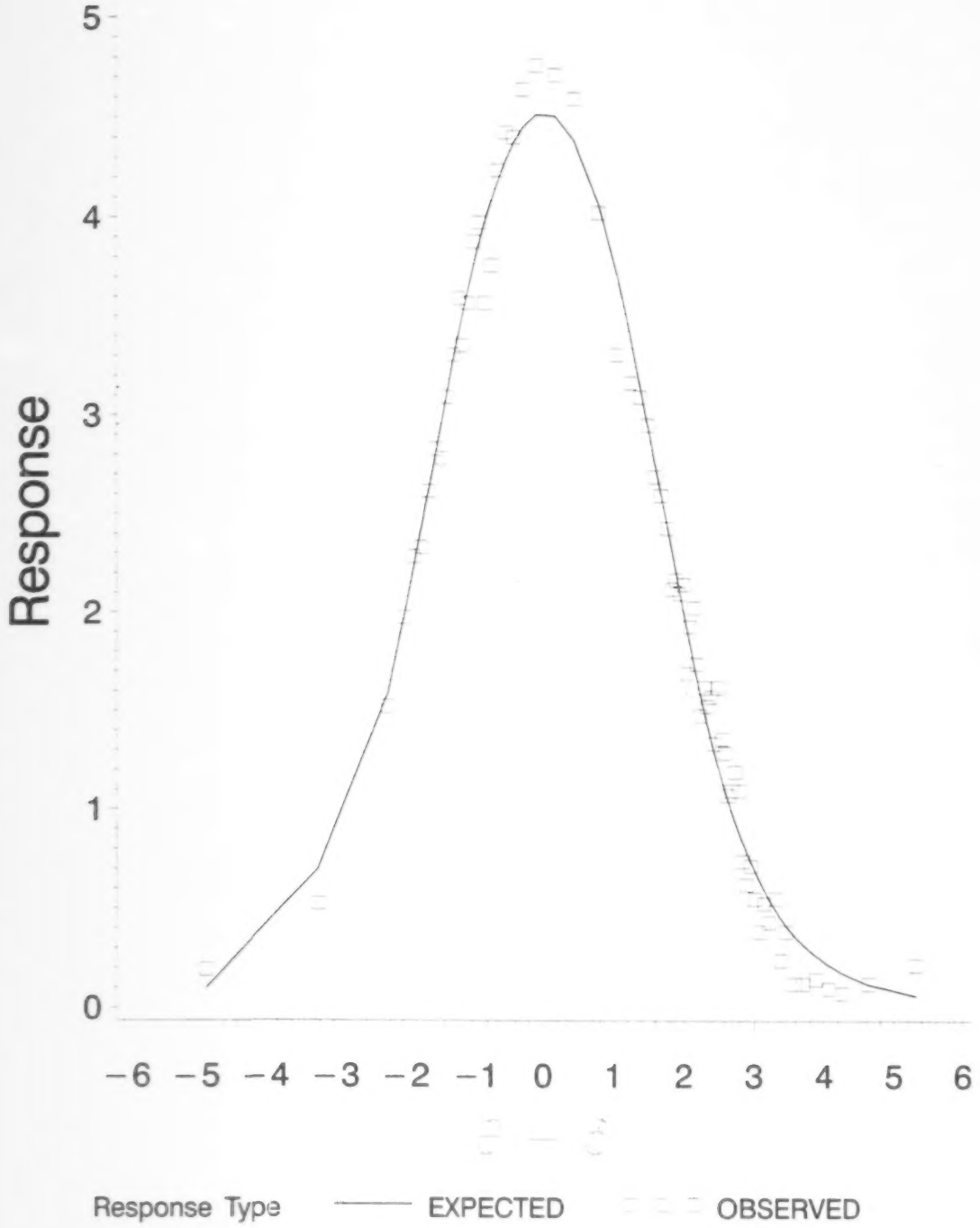
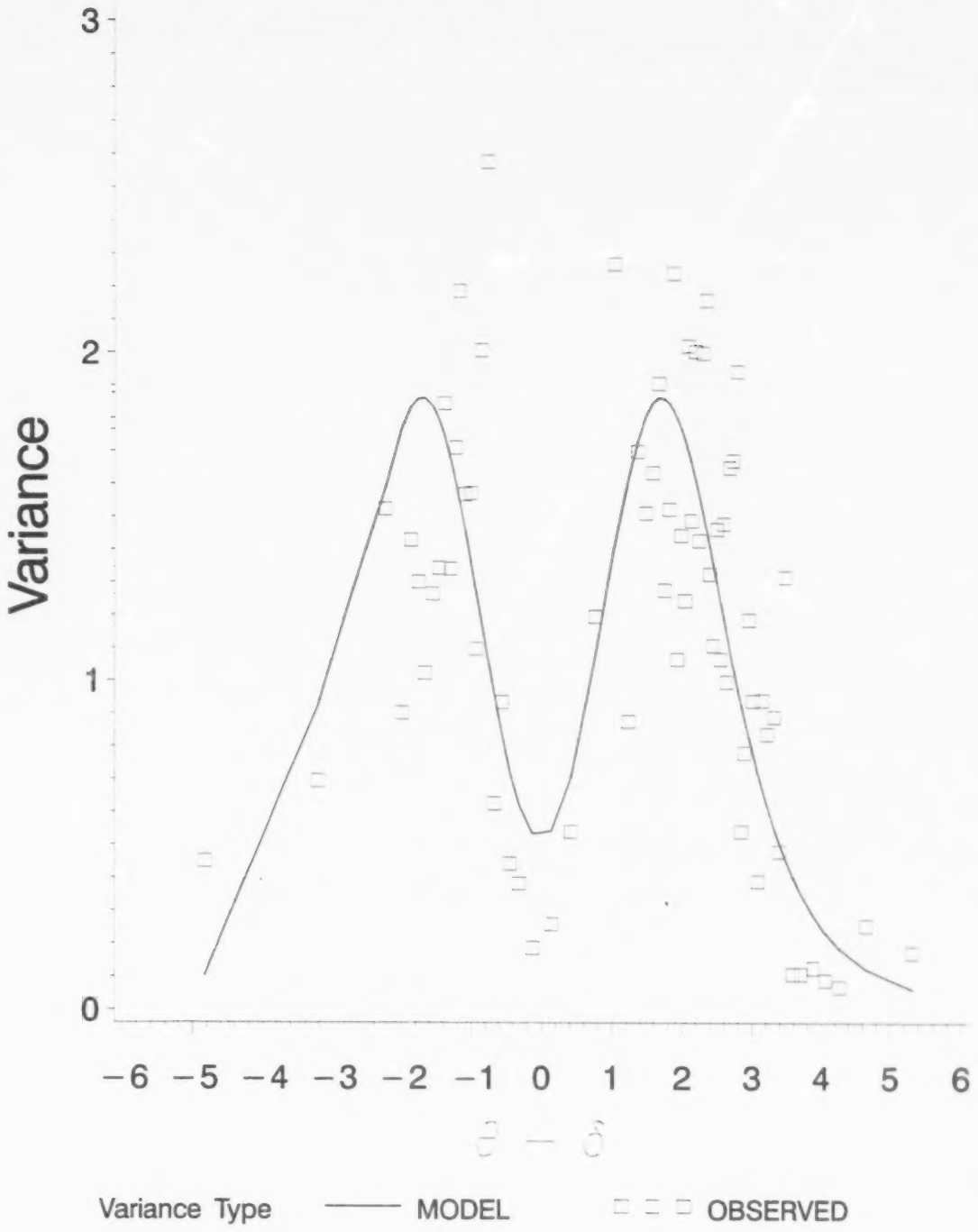

Response
5
4
3
2
1
0
−6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6
Response Type
EXPECTED
OBSERVED

Figure 15


Variance
3
2
1
0
−6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6
Variance Type
MODEL
OBSERVED

Appendix A

Developing Initial Parameter Values

Initial estimates of GUM parameters are derived most easily by first dichotomizing the graded response data such that any level of agreement is scored as 1 and all other responses are scored as 0. The resulting binary scores lead to less cumbersome solutions for initial estimates, and these simpler solutions provide adequate input to the algorithm. For binary scores, there is only one τ_j parameter that must be estimated in the GUM, and we will denote this parameter as τ_b. The initial estimate of τ_b is obtained by first setting all θ_n equal to 0, and then setting the δ_i value of the most frequently endorsed item to 0. Consequently, the estimate of τ_b must be large enough to account for the proportion of endorsements observed for the most frequently endorsed item. Let v denote the most frequently endorsed item, let s_v equal the number of endorsements observed for item v, and let N equal the number of subjects in the sample. Presuming that item v is located at δ_v=0 and all θ_n=0, then the expected proportion of endorsements for item v is:

$$Pr[X_{nv} = 1] = \frac{\exp(-\tau_b)+\exp(-\tau_b)}{\exp(0)+\exp(-\tau_b)+\exp(-\tau_b)+\exp(0)} = \frac{\exp(-\tau_b)}{1+\exp(-\tau_b)} \quad \textbf{(40)}$$

Now, if we set the observed proportion of endorsements for item v equal to the expected proportion, then an algebraic solution for τ_b can be obtained as follows:

$$\frac{\exp(-\tau_b)}{1+\exp(-\tau_b)} = \frac{s_v}{N}$$

$$\tau_b = -\ln\left[\frac{s_v}{N-s_v}\right]. \qquad (41)$$

This estimate of τ_b can be used to derive initial values for both δ_i and τ_j.[6] For example, if we maintain that all θ_n are equal to zero and that the location of the item with the largest proportion of endorsements is δ_v=0, then the absolute values of the remaining item locations can be obtained by solving the following equation with respect to δ_i ($i \neq v$):

$$Pr[X_{ni} = 1] - \frac{s_i}{N} = 0$$

$$= \frac{\exp[-\delta_i - \tau_b] + \exp[-2\delta_i - \tau_b]}{\exp[0] + \exp[-\delta_i - \tau_b] + \exp[-2\delta_i - \tau_b] + \exp[-3\delta_i]} - \frac{s_i}{N}. \qquad (42)$$

Equation 42 does not have a closed form and must be solved numerically for each δ_i. This is easily accomplished using a bisection algorithm. The solutions derived from equation 42 are the absolute values of the corresponding δ_i estimates. The sign of each estimate is obtained from the signs of the pattern loadings from the first principal component of the interitem correlation matrix (calculated from graded responses). Davison (1977) has shown that these signs will correspond to the direction of the item when responses exhibit an unfolding structure.

The estimate of τ_b can also be used to calculate initial values for τ_j. If one assumes that the

---

[6] It seems practical to avoid extreme estimates of τ_b when calculating initial values. Therefore, τ_b can be restricted to a reasonable range of possible values (e.g., -3.0 to -.5) simply by rescaling the observed proportion of endorsements for each item so that equation 41 yields an acceptable value. If this is done, then the same rescaled proportions should be used in equation 42.

introduction of a graded response scale will simply decrease the interthreshold distance in a linear fashion and that the distances between successive thresholds are equal, then the following initial estimate of τ_j is reasonable:

$$\tau_j = (C + 1 - j)\tau_b / k , \qquad \textbf{(43)}$$

where k is the number of observable response categories scored as 1 during the dichotomization process.

Finally, an initial estimate of each θ_n parameter can be obtained by considering those items that a given individual has endorsed to at least some extent. A weighted average of the initial δ_i estimates associated with such items can be obtained using the graded response scores as the weights, and this average provides a suitable initial estimate of θ_n.

Appendix B

Empirical item characteristic curves were developed for the 24 items from the capital punishment questionnaire. The curves were generated using two distinct sets of data and without any reliance on the GUM. The first set of data consisted of favorability ratings for each statement obtained from an independent sample of 117 University of South Carolina undergraduates. The ratings were on a 9-point scale with larger values indicating more favorable sentiment toward capital punishment. The favorability ratings were averaged for each statement, and the resulting averages served as rough estimates of statement locations on the latent attitude continuum. The graded disagree-agree responses provided by the 245 subjects from the example constituted the second set of data. These data were used to develop empirical attitude estimates. Each subject's attitude estimate was derived by taking a weighted median of the locations associated with those statements the subject endorsed to at least some extent. The median was weighted to emphasize the level of agreement exhibited by the subject. "Strongly agree" responses were weighted most heavily, followed by "agree" and "slightly agree" responses, respectively.[7]

The empirical attitude estimates were used to divide subjects into 15 relatively homogenous attitude groups with each group containing 16 or 17 subjects. The average response to a given item was then calculated within each attitude group. These data are portrayed in Figures B1 through B6. Each figure contains four item characteristic curves in which the average response (ordinate) is portrayed as a function of the mean attitude score within the attitude group (abscissa). The item content is listed above its characteristic curve along with its mean

---

[7] The median was weighted by manipulating the frequency count of statement locations within the distribution of endorsed statements. The frequencies were set equal to 3, 2, and 1 for "strongly agree", "agree" and "slightly agree" statements, respectively.

favorability rating. The 24 items shown in Figures B1-B6 have been ordered on the basis of their mean favorability rating.

Insert Figures B1-B6 About Here

The characteristic curves associated with items 11, 13 and 20 indicated that these items were relatively independent of attitudes toward capital punishment. These items generally elicited disagreement from all subjects regardless of their attitudes. In contrast, the characteristic curves for the remaining 21 items were more or less consistent with the hypothesis that responses resulted from an ideal point process. Statements that were clearly unfavorable with regard to capital punishment (i.e., statements 1-10) exhibited response patterns that were negatively related to attitude. Subjects with negative attitudes toward capital punishment endorsed these statements to a higher degree than those with moderate or positive attitudes, respectively. Conversely, those statements that expressed clearly positive sentiment (statements 19, 21-24) showed the opposite response pattern. They were endorsed most by subjects with positive attitudes toward capital punishment and were subscribed to less frequently by subjects with moderate or negative attitudes. Most importantly, the statements which expressed more moderate positions toward capital punishment (statements 12, 14-18) were endorsed most by individuals with relatively moderate attitudes. Subjects with more extreme opinions, whether positive or negative, generally endorsed these items less frequently.

Additional Figure Captions For Appendix B

Figure B1. Empirical item characteristic curves for statements 1 through 4 of the capital punishment scale.

Figure B2. Empirical item characteristic curves for statements 5 through 8 of the capital punishment scale.

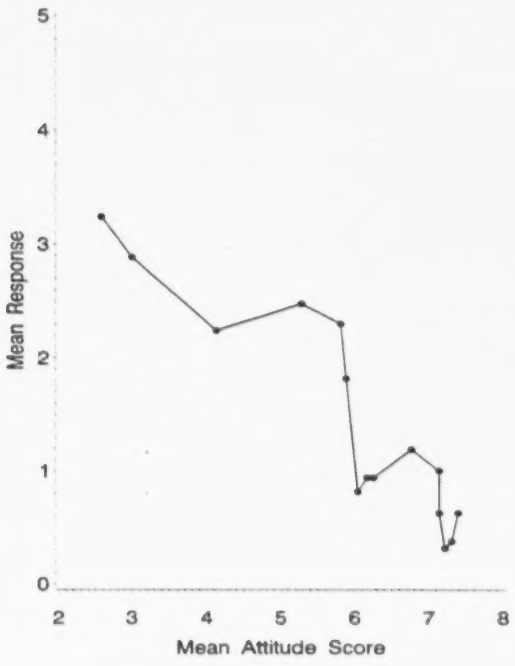
Figure B3. Empirical item characteristic curves for statements 9 through 12 of the capital punishment scale.

Figure B4. Empirical item characteristic curves for statements 13 through 16 of the capital punishment scale.

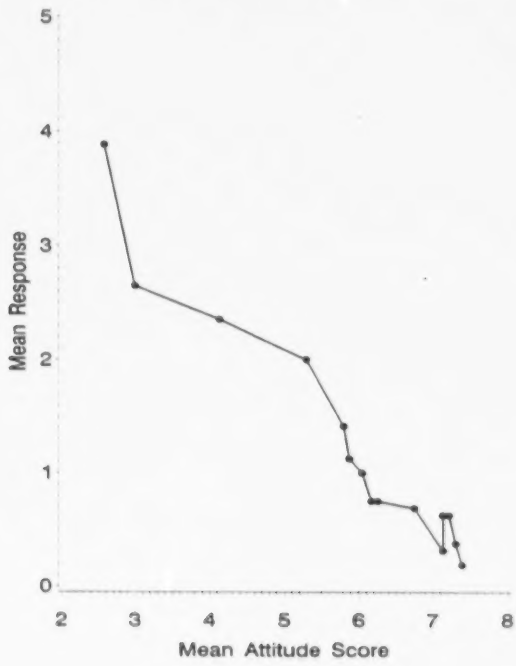
Figure B5. Empirical item characteristic curves for statements 17 through 20 of the capital punishment scale.

Figure B6. Empirical item characteristic curves for statements 21 through 24 of the capital punishment scale.

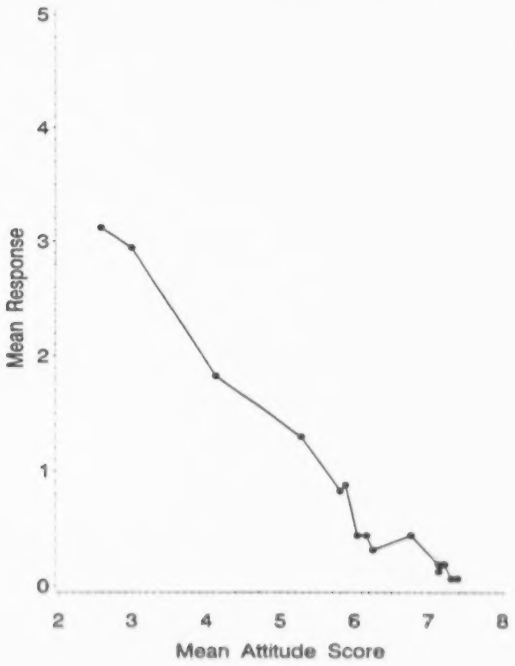
1. Capital punishment is the most hideous practice of our time. (2.0)



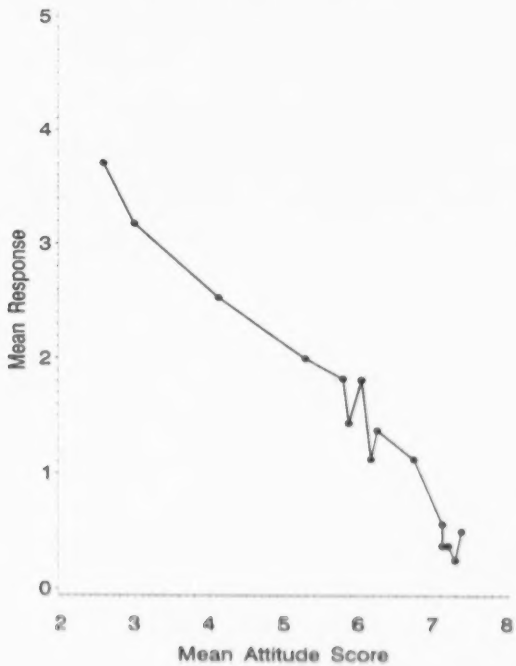
2. Capital punishment is absolutely never justified. (2.3)



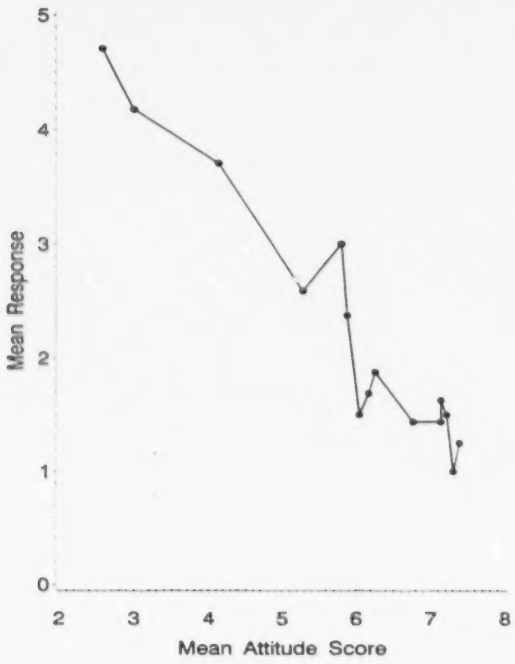
3. I do not believe in capital punishment under any circumstances. (2.4)



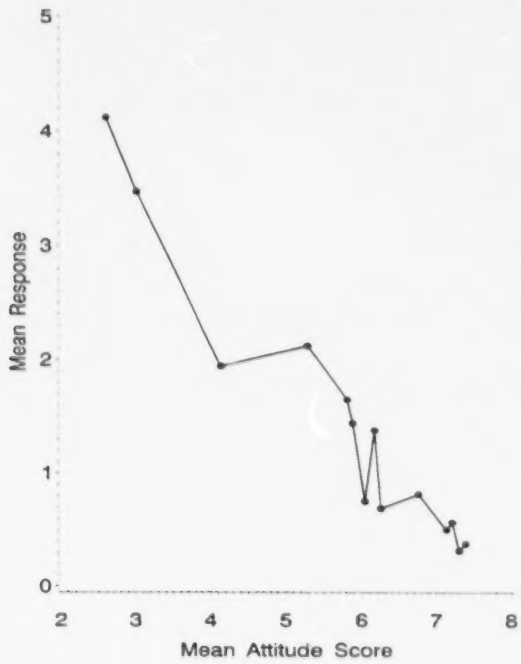
4. Execution of criminals is a disgrace to our society. (2.4)



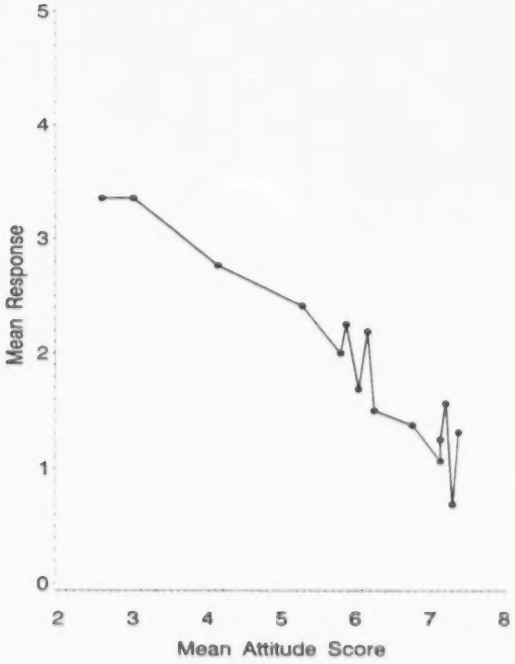
5. The state cannot teach the sacredness of human life by destroying it. (2.4)



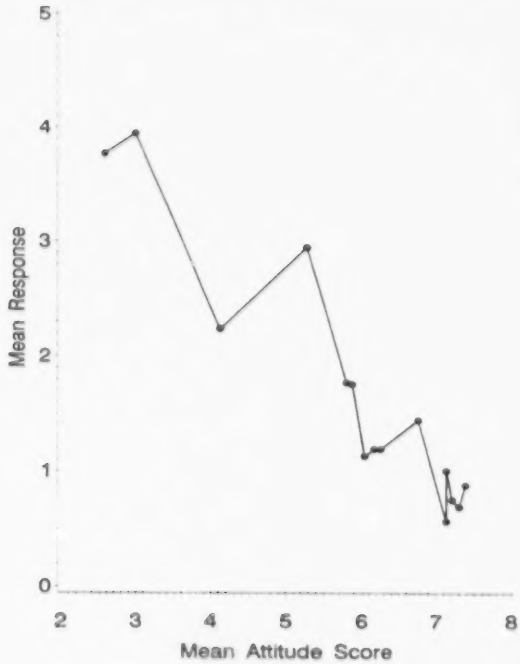
6. We cant't call ourselves civilized as long as we have capital punishment. (2.6)



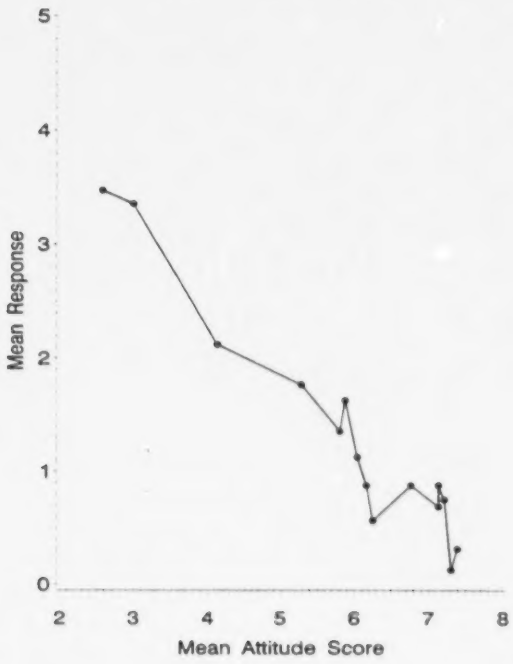
7. Capital punishment has never been effective in preventing crime. (2.9)



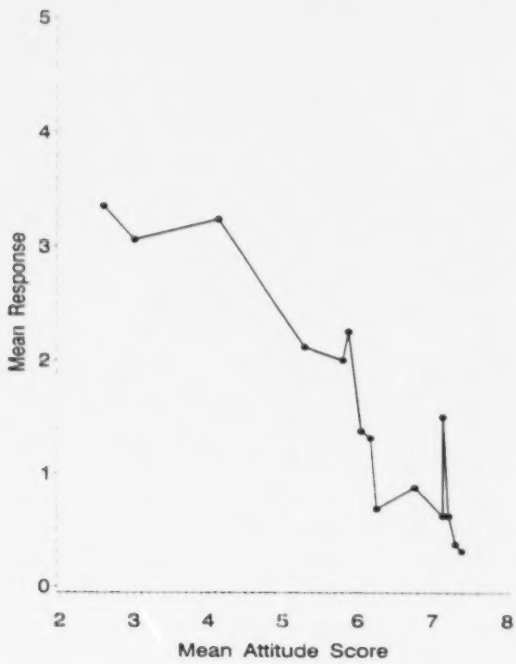
8. Capital punishment cannot be regarded as a sane method of dealing with crime. (3.0)



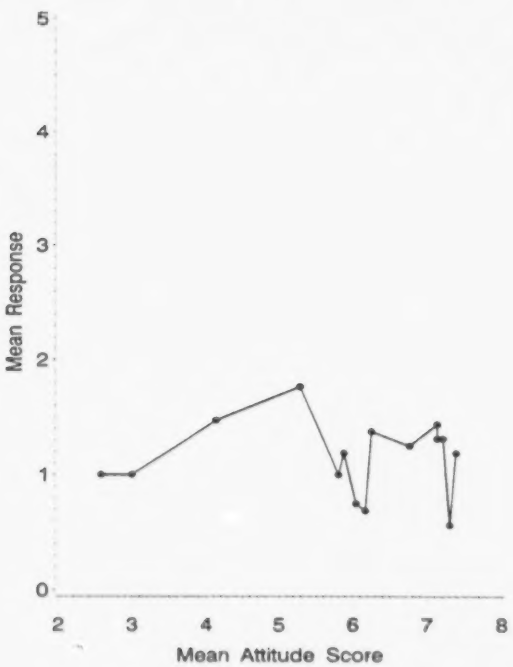
9. Capital punishment is not necessary in modern civilization. (3.0)



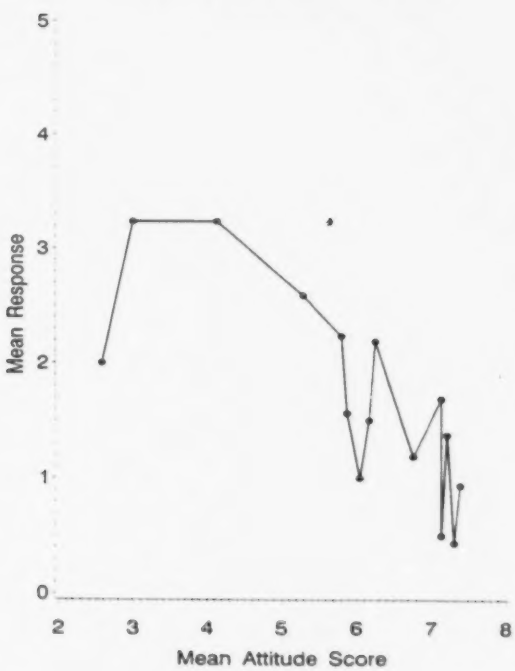
10. Life imprisonment is more effective than capital punishment. (3.1)



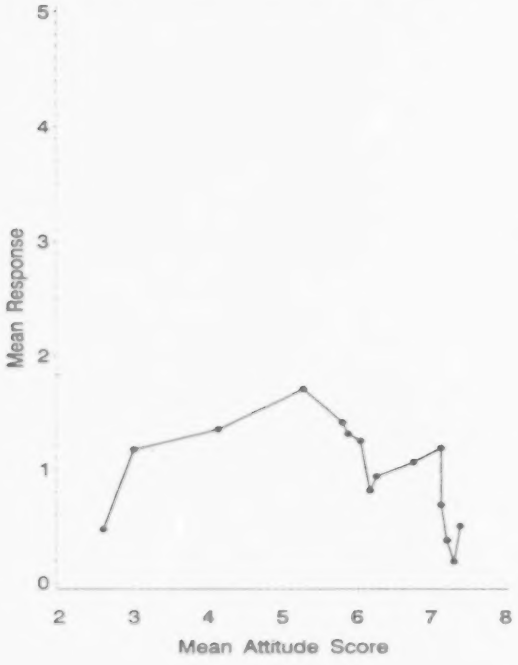
11. I think the return of the whipping post would be more effective than capital punishment. (3.6)



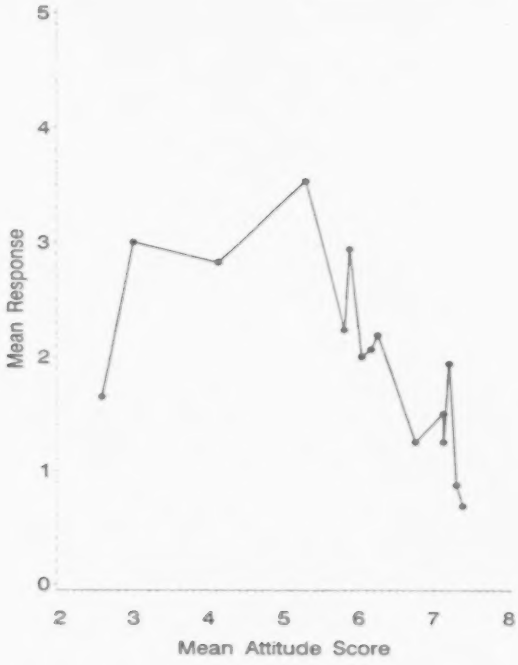
12. I don't believe in capital punishment but I'm not sure it isn't necessary. (4.8)



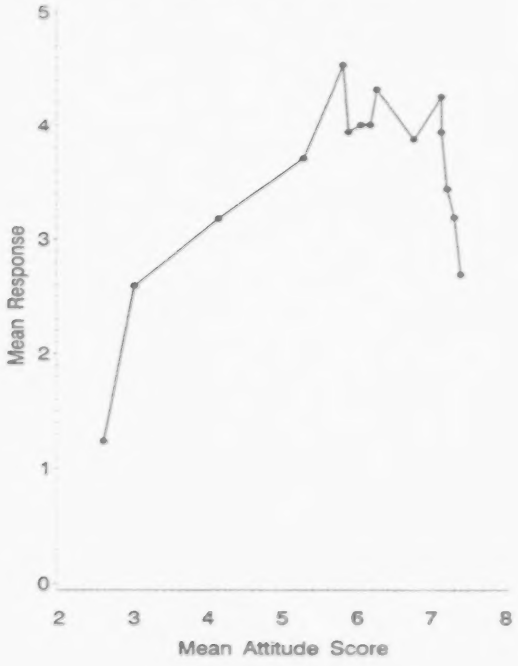
13. It doesn't make any difference to me whether we have capital punishment or not. (4.9)



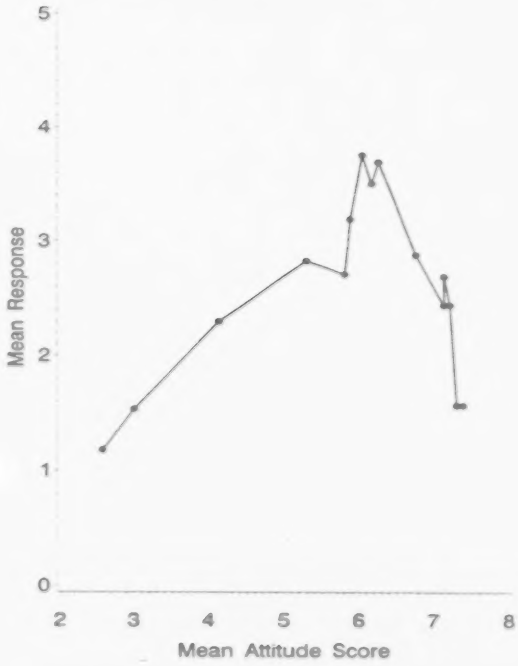
14. I do not believe in capital punishment but it is not practically advisable to abolish it. (5.1)



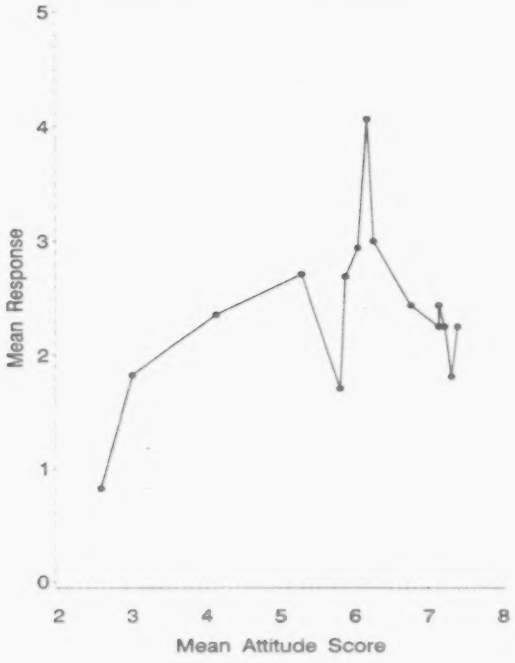
15. I think capital punishment is necessary but I wish it were not. (5.8)



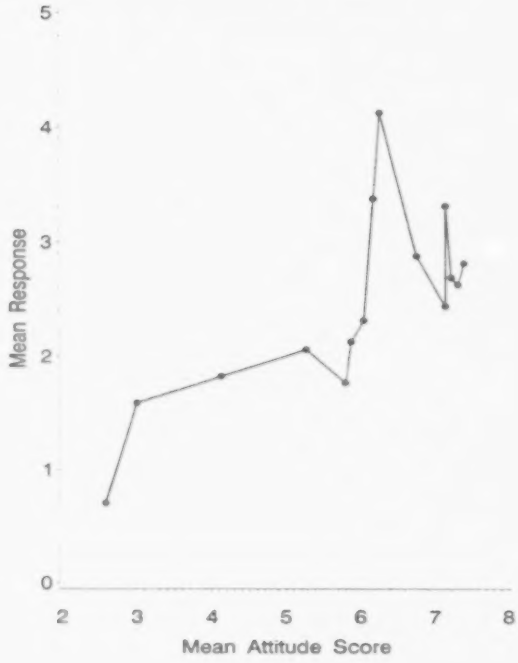
16. Capital punishment is wrong but is necesssary in our imperfect civilization. (5.9)



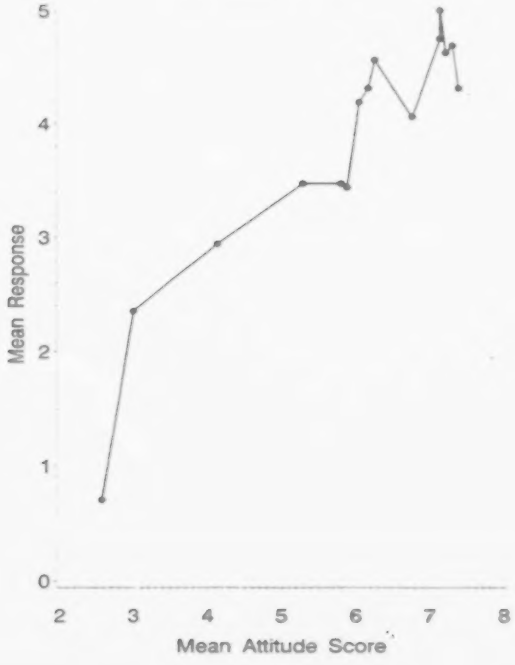
17. Capital punishment is justified only for premeditated murder. (6.1)



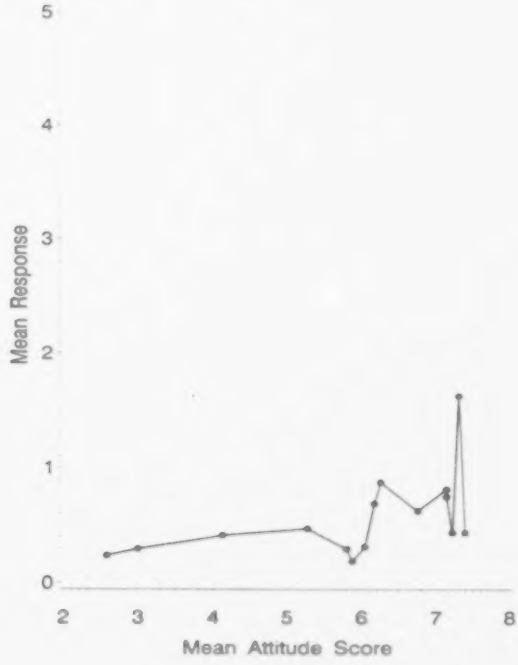
18. Capital punishment may be wrong but it is the best preventative to crime. (6.3)



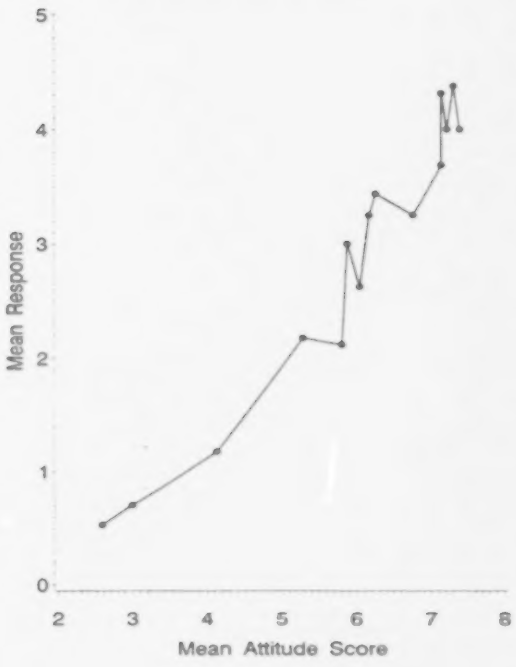
19. We must have capital punishment for some crimes. (7.1)



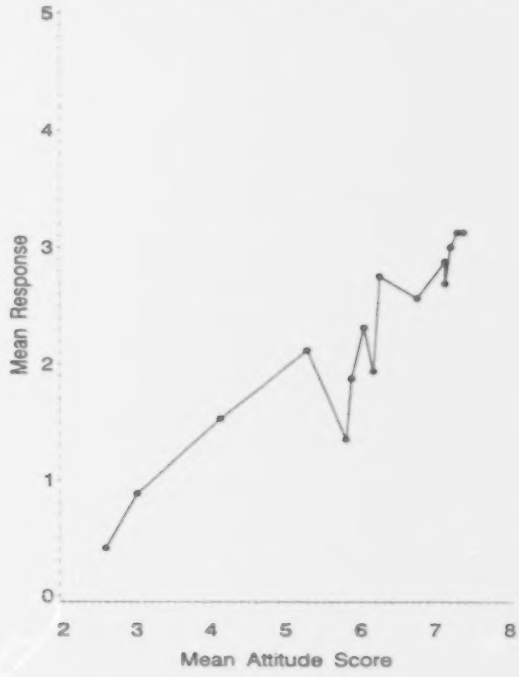
20. Every ciminal should be executed. (7.2)



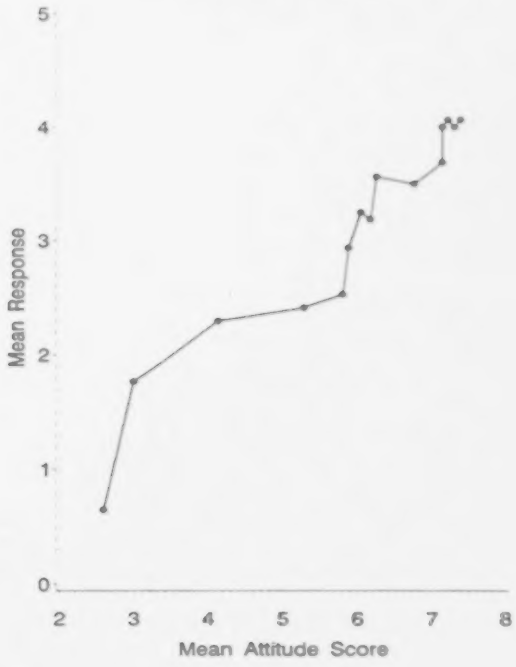
21. Capital punishment should be used more often than it is. (7.3)



22. Any person, man or woman, young or old, who commits murder, should pay with his own life. (7.4)



23. Capital punishment gives the criminal what he deserves. (7.7)



24. Capital punishment is just and necessary. (7.9)